# GC-MS Analysis of Non-volatile Small Molecules and Metabolomics Workflows

**Adam Heuberger, Ph.D.**

**Assistant Professor**

**Colorado State University**

**adam.Heuberger@colostate.edu**

# Outline

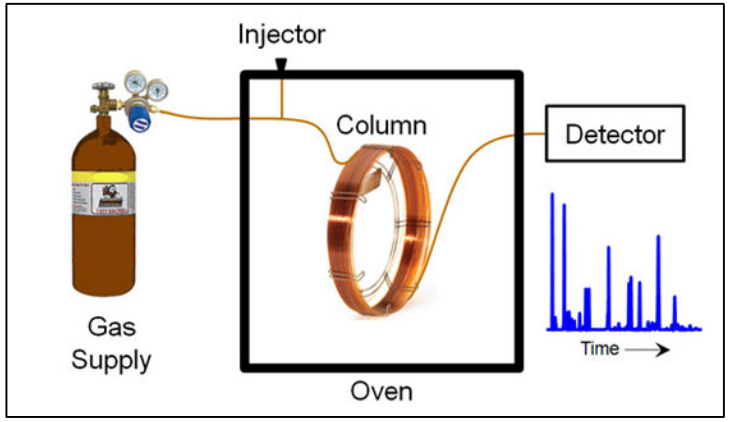
(1) How to analyze non-volatile small molecules using GC-MS

(2) GC-MS metabolomics workflows including data processing

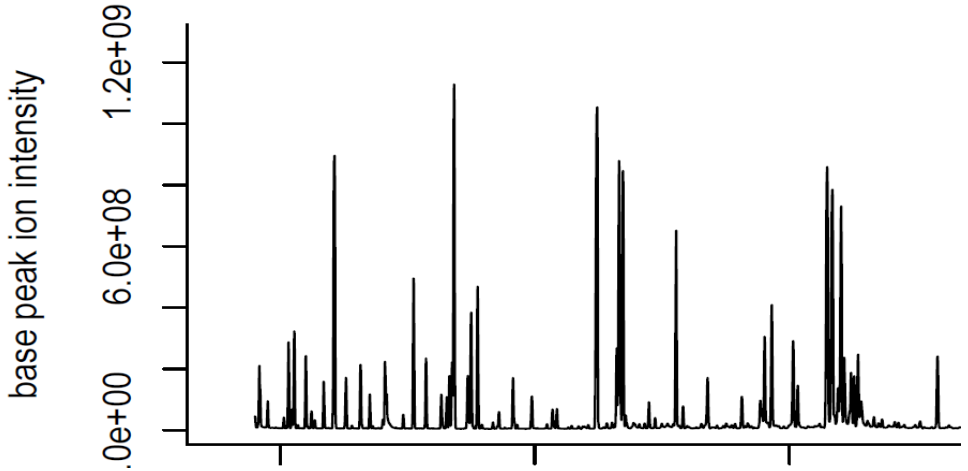
(3) Statistics used to interpret GC-MS metabolomics data

# Gas Chromatography Provides Great Detection For Small Molecules

**separation**




**profile (time & m/z)**



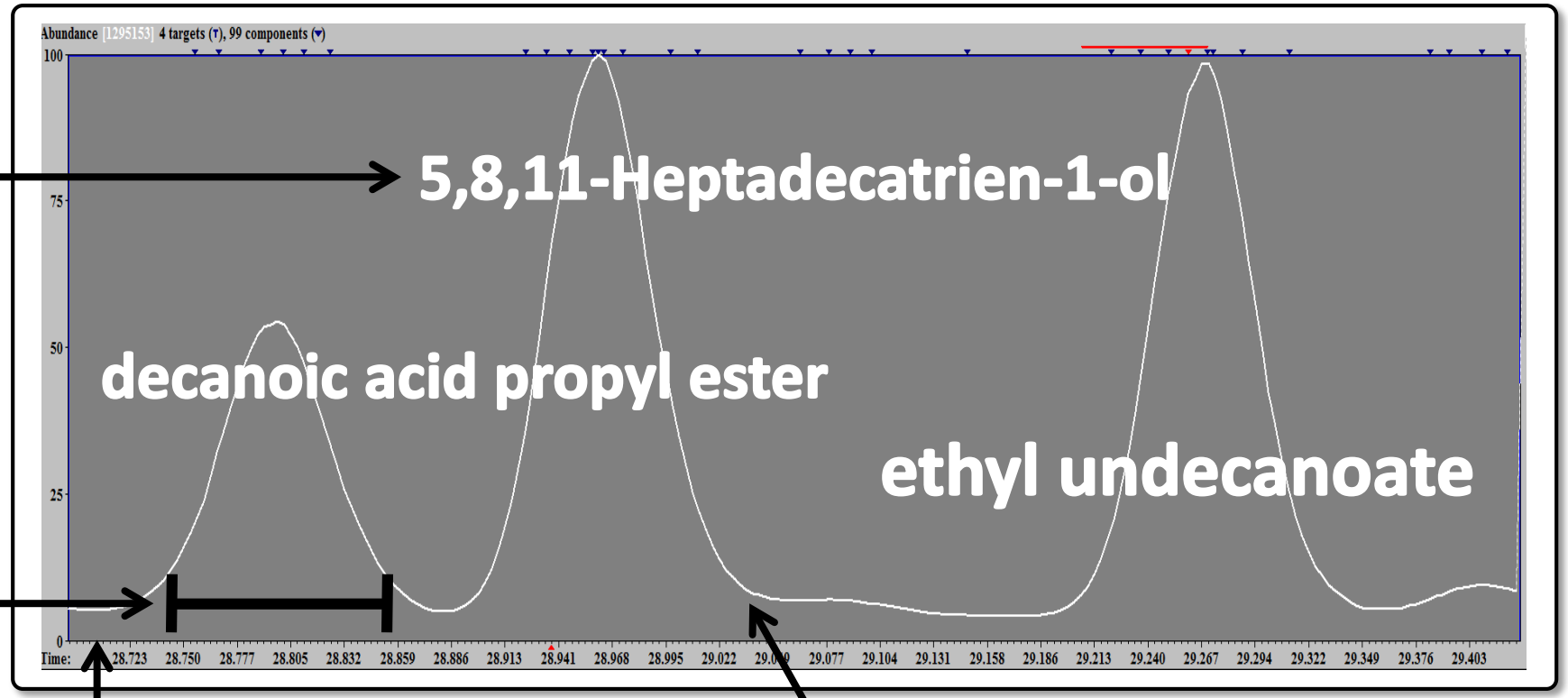
# For Small Molecules, GC-MS is Selective, Sensitive, and Provides Excellent Quantitation

**descriptive (MS)**

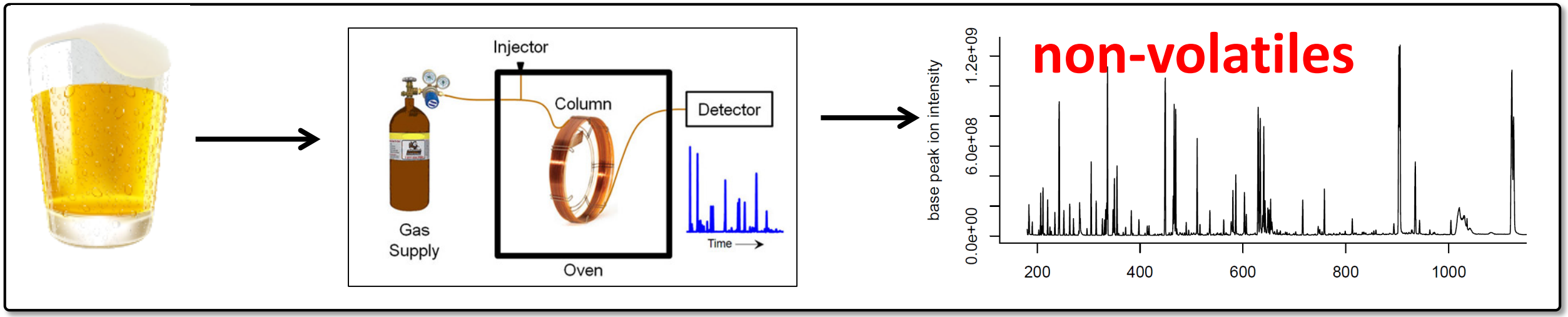
**narrow peaks (~6 sec)**

**low noise**

**low tailing**



# There Is Value in Using GC-MS for Non-volatile Small Molecules




## *robust*

- **high-throughput**
- **reproducible**
- **sensitive**

## *resourced*

- **standard methods**
- **retention index and MS databases**

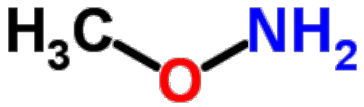
## *reduced*

- **cost-effective**

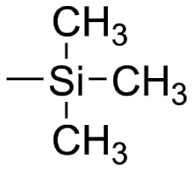
# Non-volatiles Are Amenable With GC-MS Analysis Via Derivatization

**Derivatization = a chemical alteration of the initial metabolite**

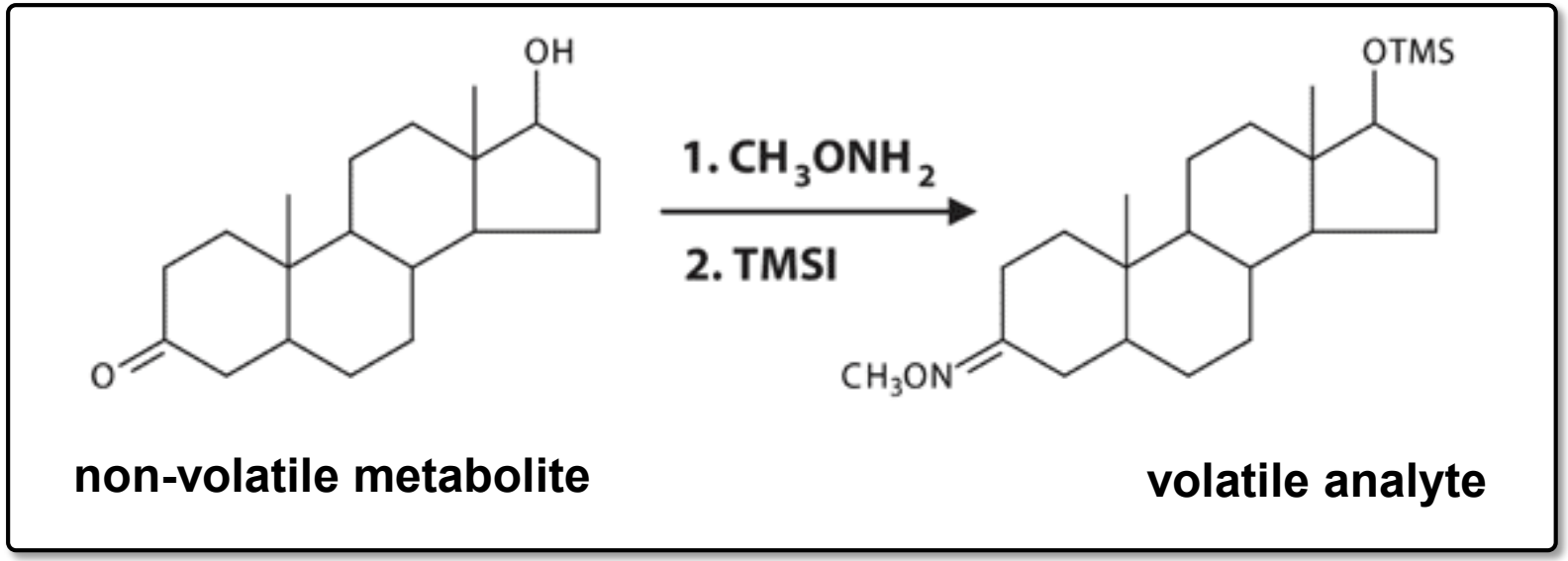
- **methoxymation**




- **silylation**



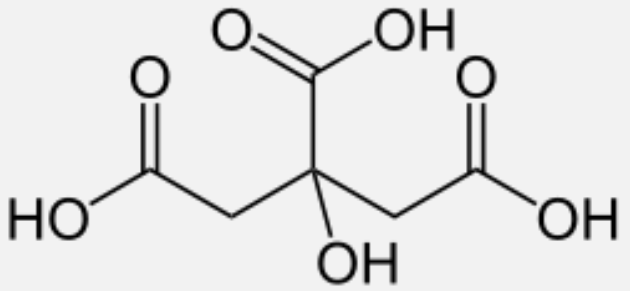

  - **-OH groups**
  - **-NH groups**



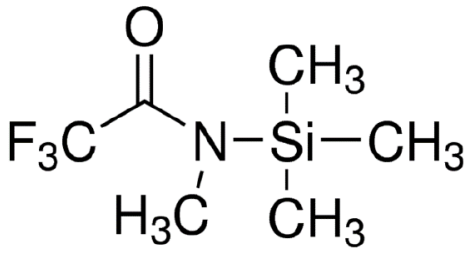
# MSTFA Is The Most Common Derivatization Reagent

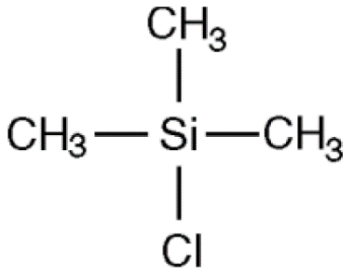
***non-volatile metabolite***

**citric acid**

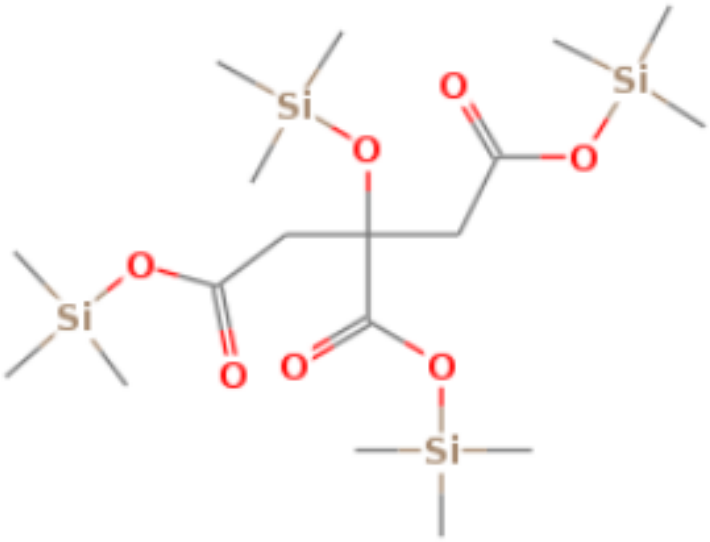



**MSTFA + 1% TMCS**



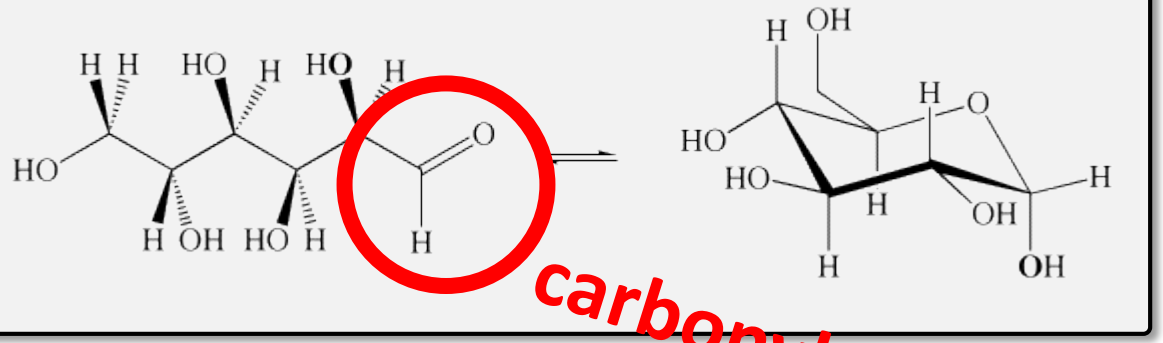

***volatile analyte***

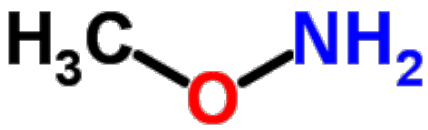



**citric acid 4TMS**

# Methoximation Is A Critical First Step For Samples With Saccharides

**glucose**

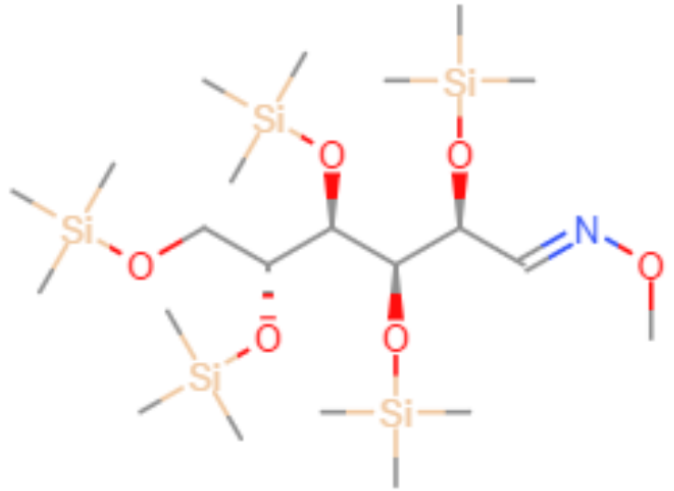


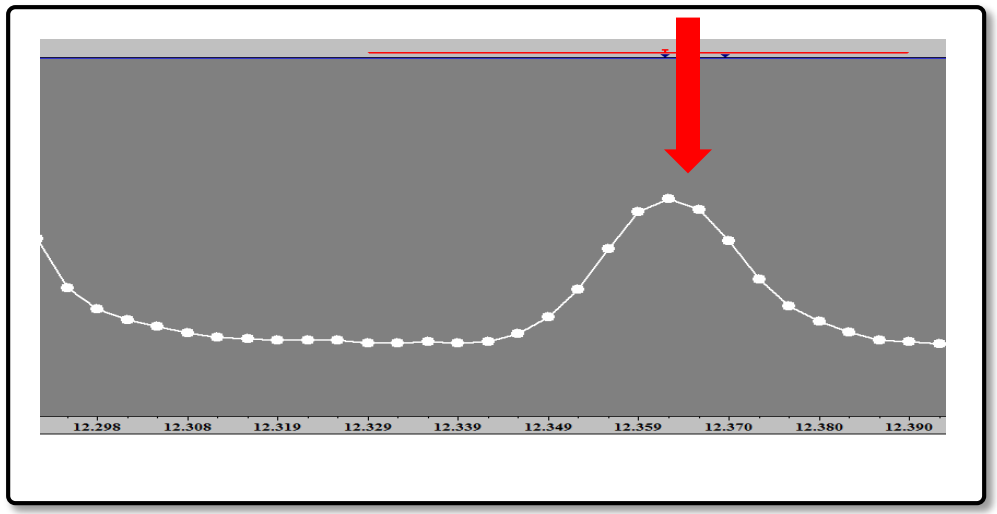

***(1) methoxyamine HCl***

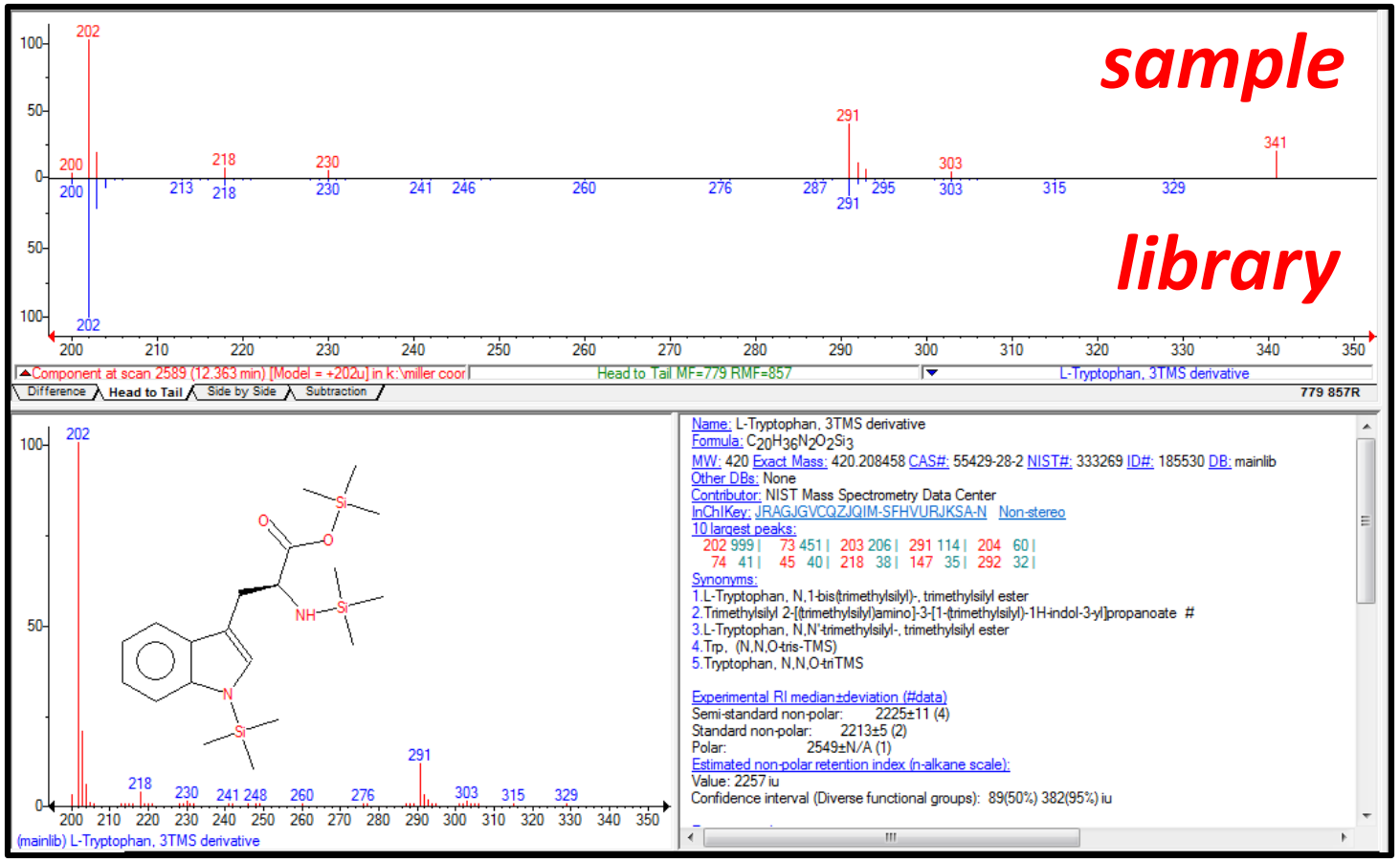
***(2) MSTFA + 1% TMCS***

**glucose 1MEOX 5TMS**



# Metabolite Databases Are Curated With Analytes



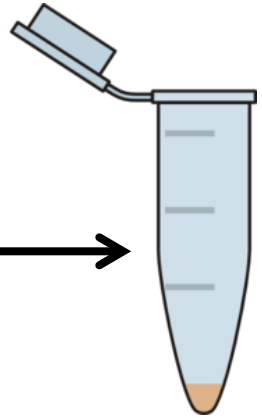

## Key databases:

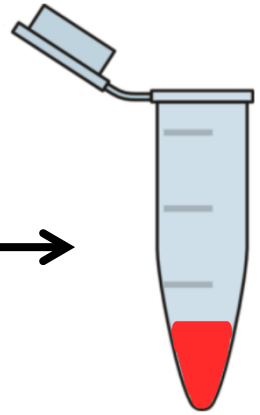
- **NIST MS**
- **Golm Metabolome Database**

# An Example Derivatization Protocol

20 µL beer

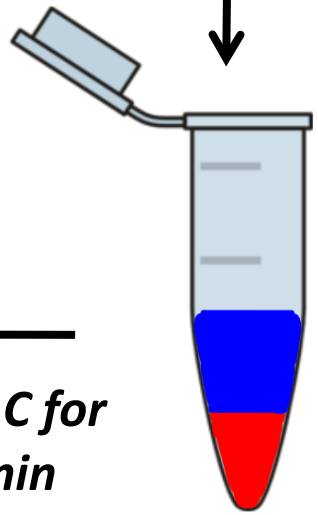
*dry*





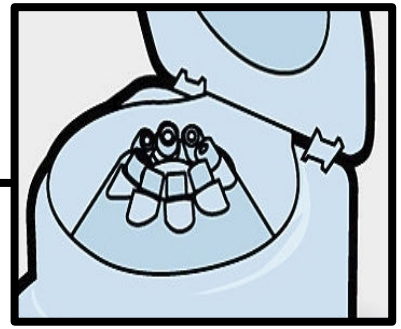
**add 50 µL 25 mg/mL methoxyamine HCl in pyridine**

*60 °C for 1.5 hrs*



**add 25 µL MSTFA + 1% TMCS**

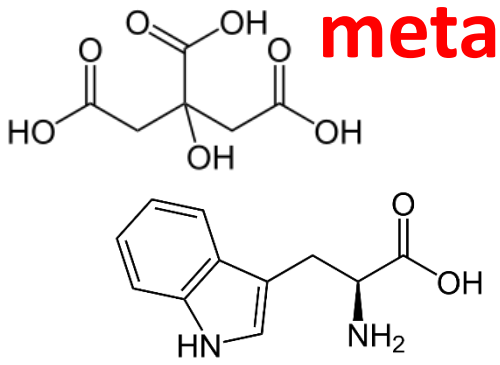
*60 ° C for 30 min*



spin 5 min

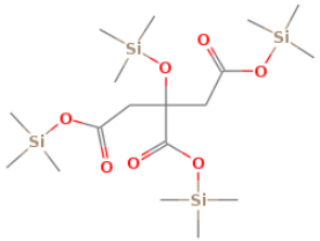
**50 µL to glass vial**

# GC-MS Non-Volatile Metabolite Profile of Beer

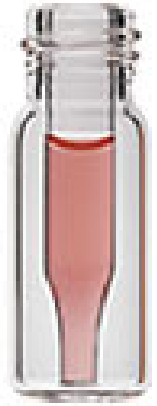


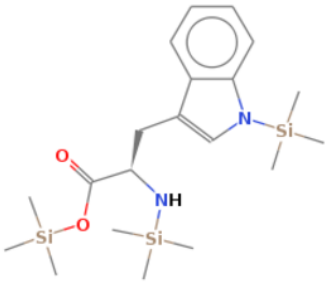

**metabolites**

**derivatize**

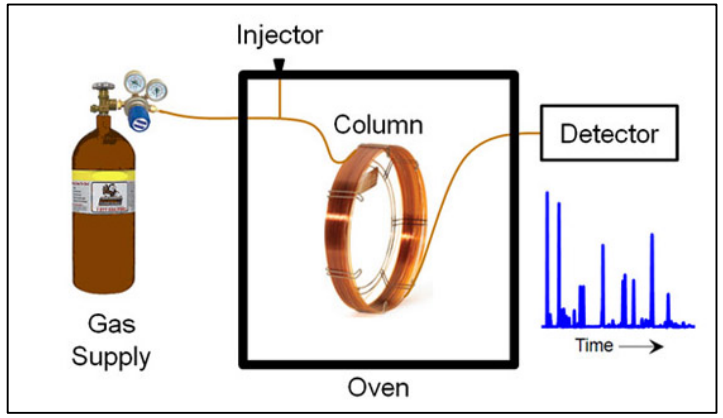


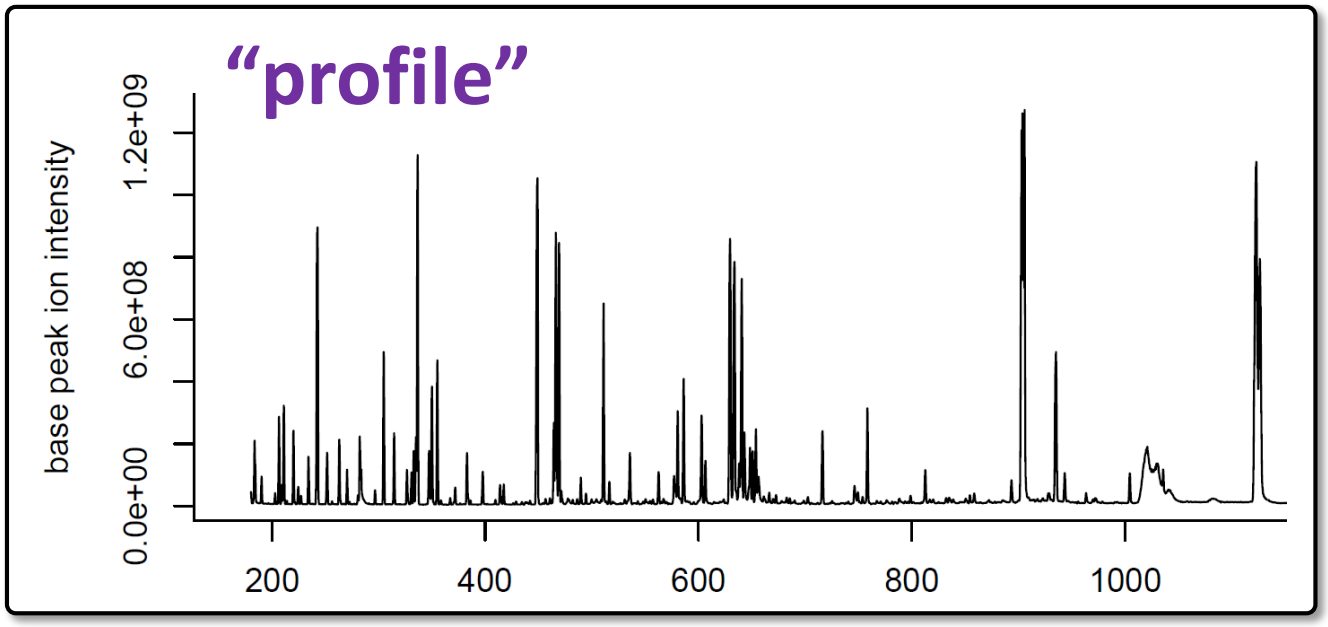
**analytes**



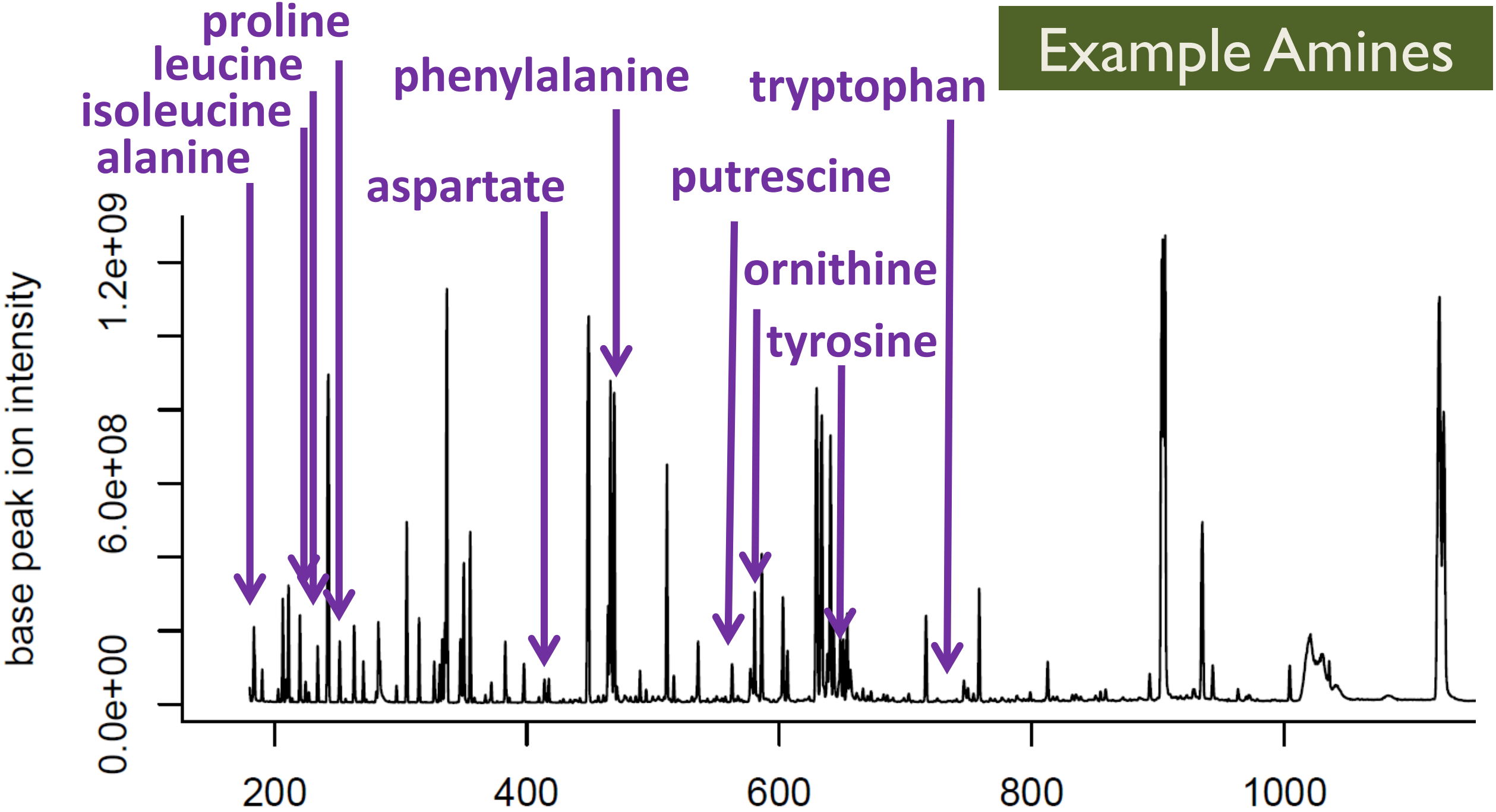



**inject 1 µL**

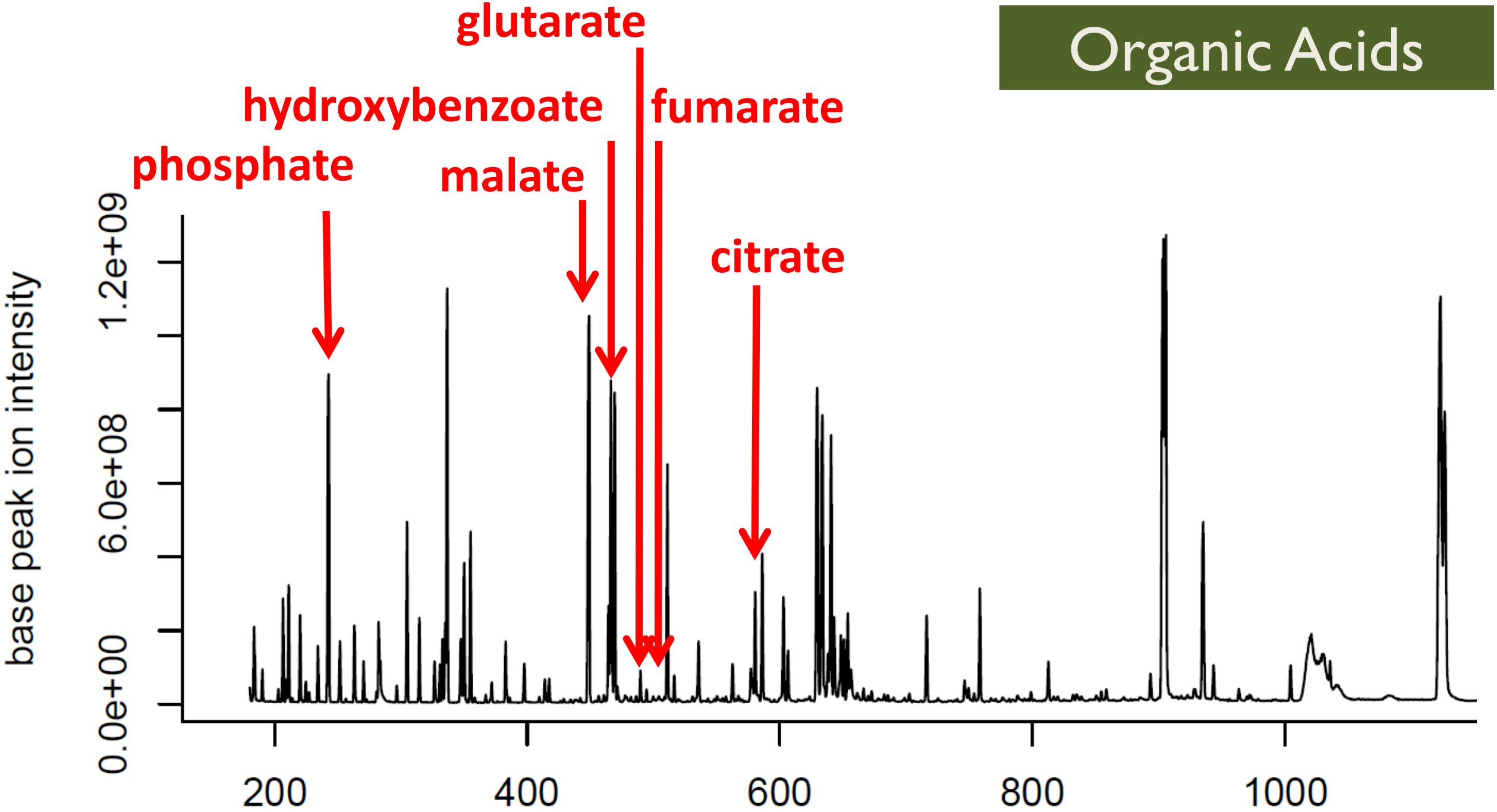


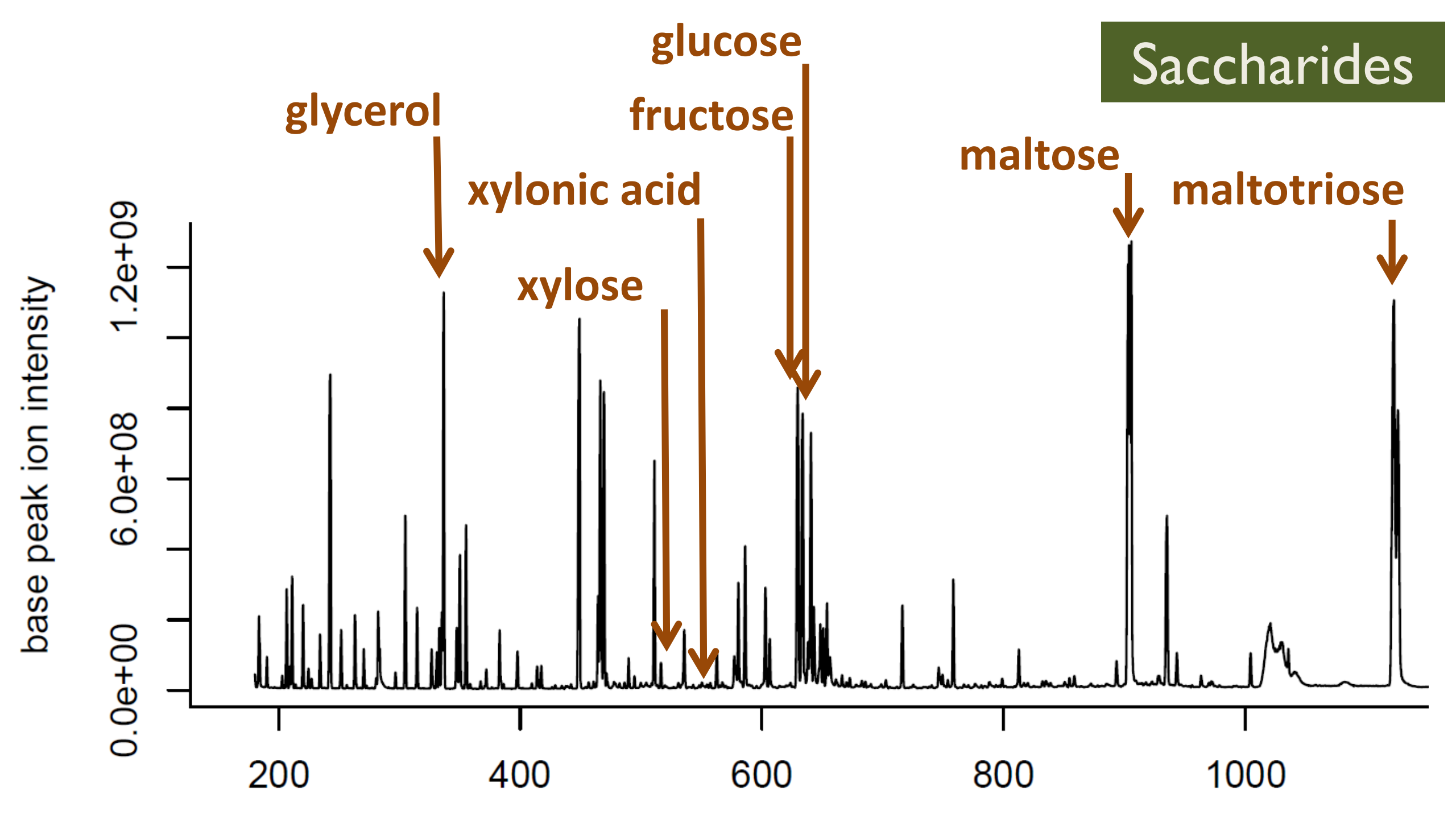



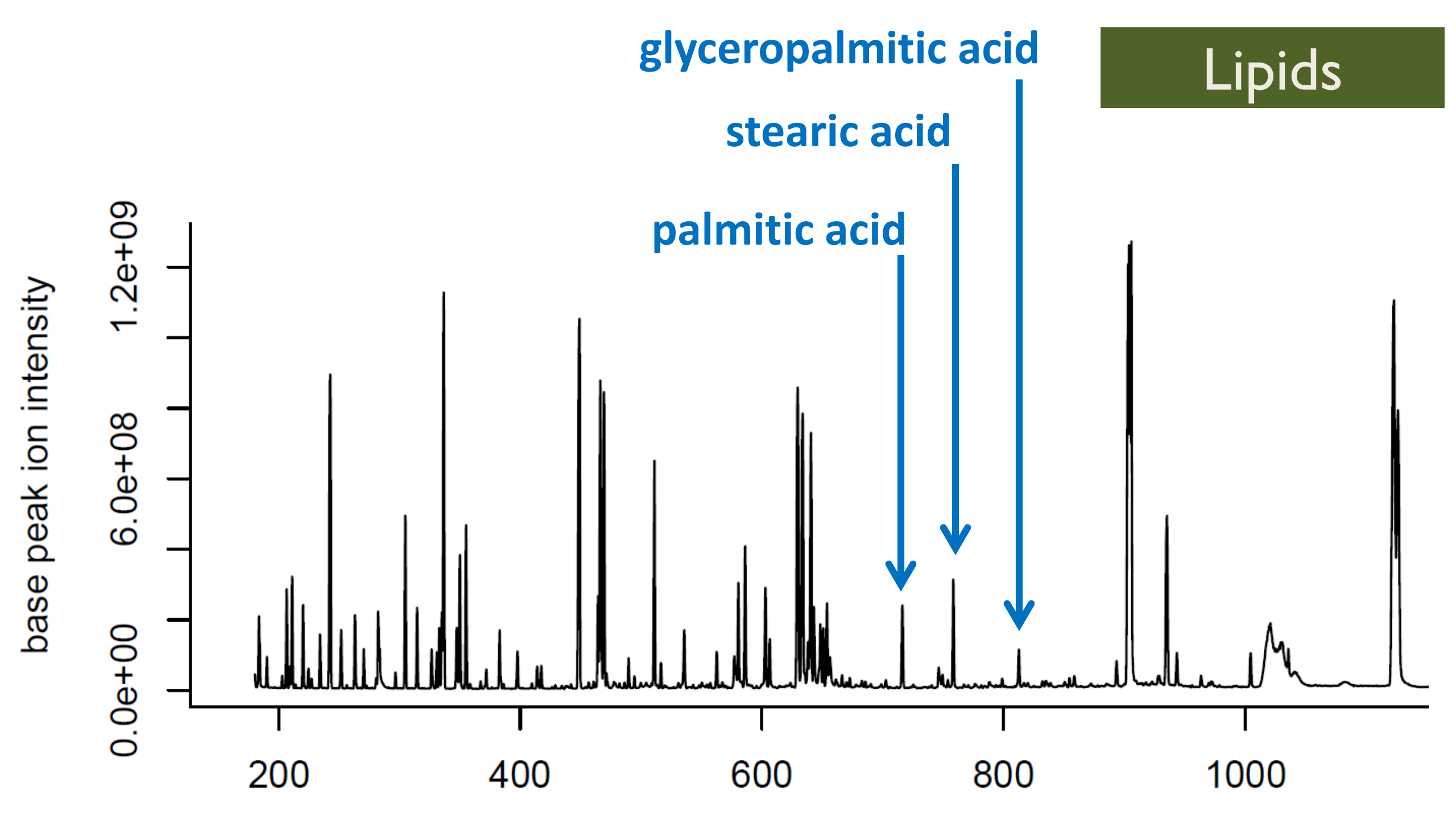
# Example Amines



# Organic Acids




Saccharides
glycerol
xylonic acid
xylose
glucose
fructose
maltose
maltotriose
base peak ion intensity
0.0e+00
6.0e+08
1.2e+09
200
400
600
800
1000


Lipids
glyceropalmitic acid
stearic acid
palmitic acid
base peak ion intensity
0.0e+00
6.0e+08
1.2e+09
200
400
600
800
1000

# Summary of Analyzing Non-volatiles via GC-MS

## THE REALLY GOOD

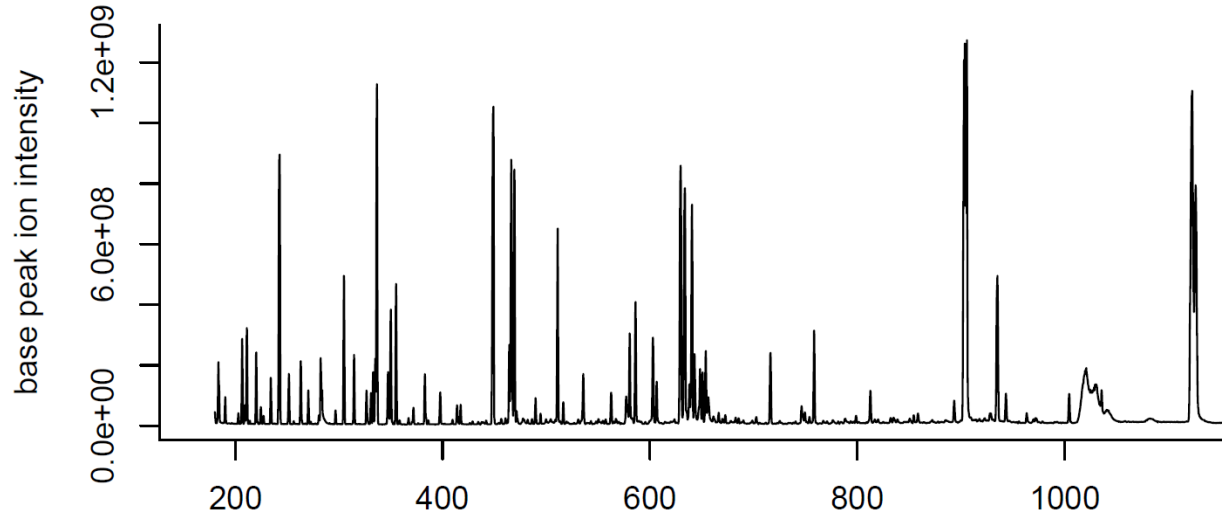
- **sensitivity**
- **selectivity – complexity is reduced due to bias for volatility**
- **reproducibility – easy to compare GC-MS studies**
- **cost -- GC systems are robust and easy to maintain**
- **complement to reverse-phase LC methods**

## THE NOT SO GOOD

- **selectivity – volatility is required for detection**
- **dirty inlets (due to oligomers, syrups) and ion sources**
- **derivatization requires time, supplies, and reagents**
- **incomplete derivatization and artifacts**

# What Makes A GC-MS Analysis 'Metabolomics?'

***Metabolomics is a comparative analysis of many metabolite profiles***




**x 5 beer types**
**x 5 process conditions**
**x 5 time points**
**= 125 samples!**

# Early Metabolomics Studies Used MSTFA + GC-MS





## Metabolite profiling for plant functional genomics

Oliver Fiehn[1]*, Joachim Kopka[2], Peter Dörmann[1], Thomas Altmann[1], Richard N. Trethewey[2], and Lothar Willmitzer[1]

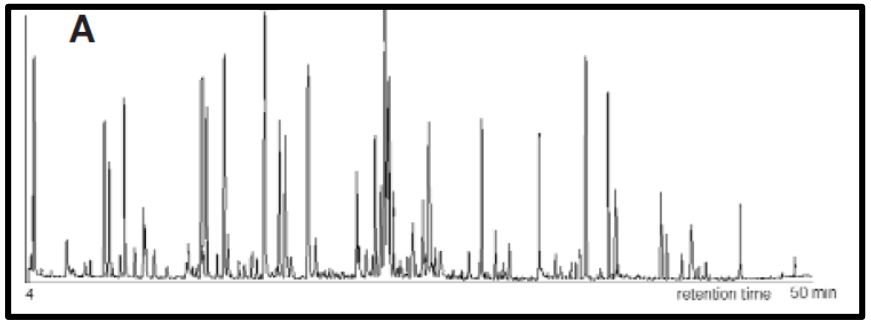
[1]*Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany.* [2]*Metanomics GmbH & Co KGaA, Tegeler Weg 33, 10589 Berlin, Germany.*
**Corresponding author (fiehn@mpimp-golm.mpg.de).*

Received 17 April 2000; accepted 21 August 2000

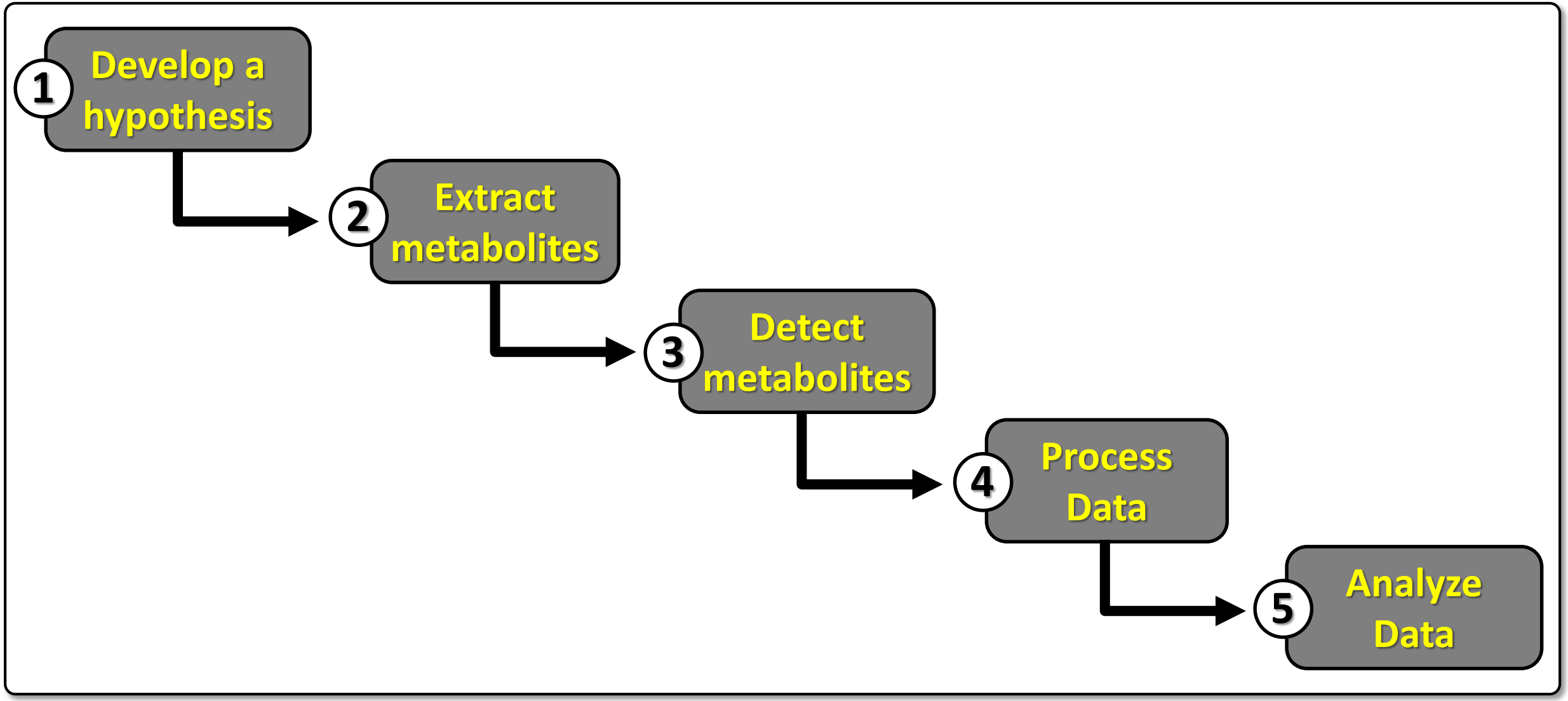
**Multiparallel analyses of mRNA and proteins are central to today's functional genomics initiatives. We describe here the use of metabolite profiling as a new tool for a comparative display of gene function. It has the potential not only to provide deeper insight into complex regulatory processes but also to determine phenotype directly. Using gas chromatography/mass spectrometry (GC/MS), we automatically quantified 326 distinct compounds from *Arabidopsis thaliana* leaf extracts. It was possible to assign a chemical structure to approximately half of these compounds. Comparison of four *Arabidopsis* genotypes (two homozygous ecotypes and a mutant of each ecotype) showed that each genotype possesses a distinct metabolic profile. Data mining tools such as principal component analysis enabled the assignment of "metabolic phenotypes" using these large data sets. The metabolic phenotypes of the two ecotypes were more divergent than were the metabolic phenotypes of the single-loci mutant and their parental ecotypes. These results demonstrate the use of metabolite profiling as a tool to significantly extend and enhance the power of existing functional genomics approaches.**

Keywords: functional genomics, *Arabidopsis thaliana*, metabolite profiling, cluster analysis, metabolomics, bioinformatics

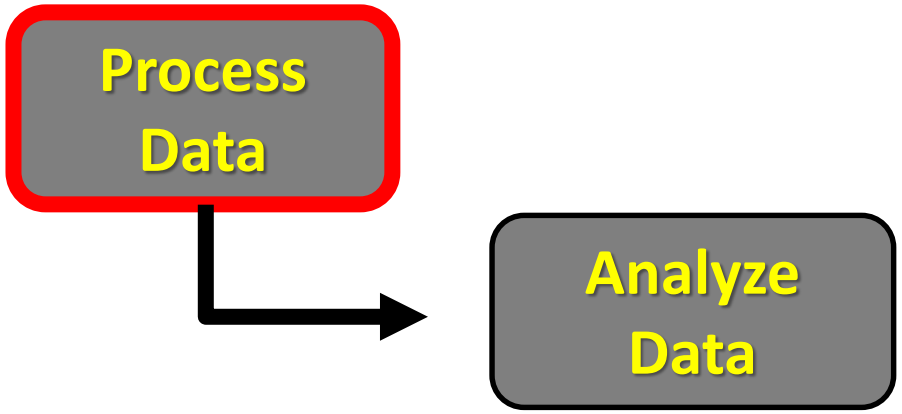
metabolite stability and volatility as reported[8]. Briefly, the lipid phase was transmethylated and trimethylsilylated for the analysis of total fatty acids, fatty alcohols, sterols, and aliphatics, whereas the polar phase was methoximated and trimethylsilylated for the analysis of hydroxy- and amino acids, sugars, sugar alcohols, organic monophosphates, (poly)amines, and aromatic acids. Metabolite sizes were in the range of ethylene glycol (62 AMU) to trisaccharides (504 AMU). Optimal reaction conditions were established as a compromise between reaction completeness



# Metabolomics Is Performed As A Workflow (5 Steps)



# MS-Metabolomics Data Processing Creates a Data Matrix of Samples and Molecular Features




**STEPS:**

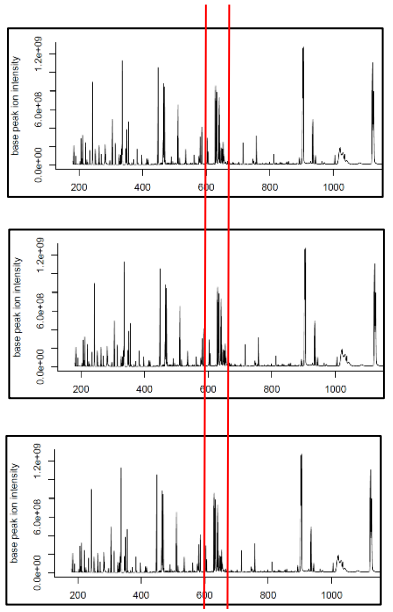
**A/B: batch peak detection/grouping**

**C: retention time alignment**

**D: intensity normalization**

**OUTPUT: data matrix**

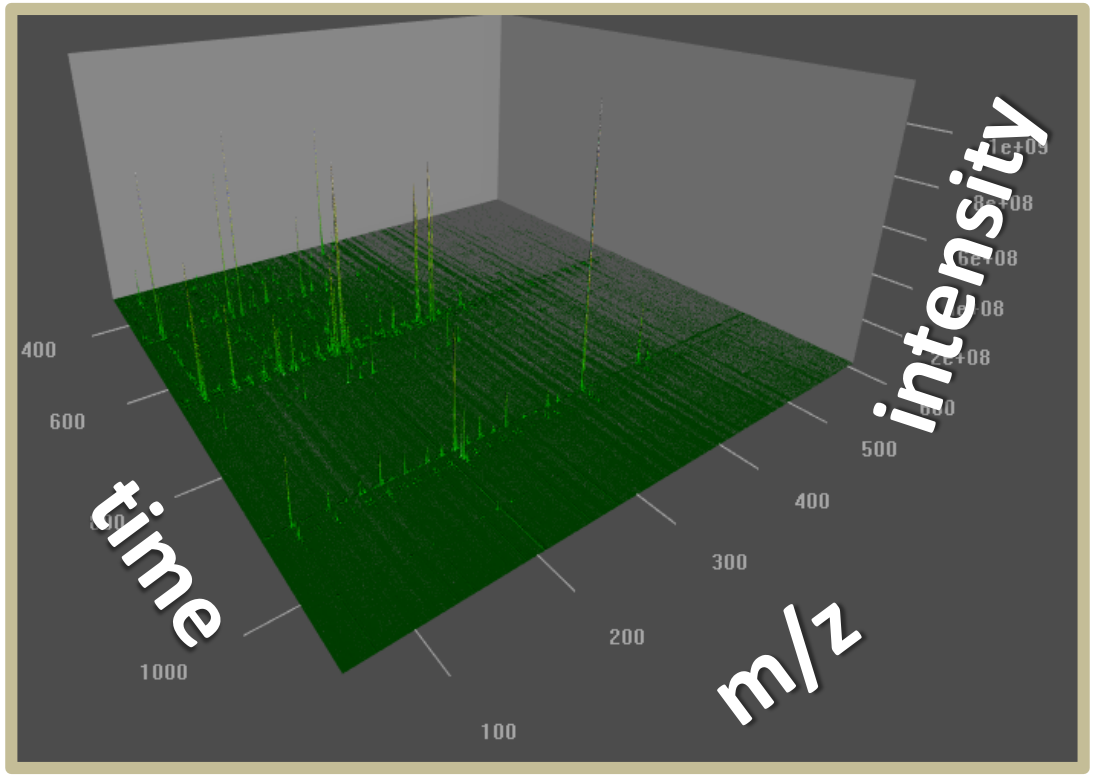
**MF = m/z_time**




| | A | B | C | D |
|---|---|---|---|---|
| 1 | sample | 76.8759_227.011 | 51.1409_227.876 | 78.1372_228.3 |
| 2 | 1 | 5049 | 15766 | 15085 |
| 3 | 2 | 1913 | 18305 | 13923 |
| 4 | 3 | 3364 | 21773 | 14991 |
| 5 | 4 | 7636 | 11918 | 9107 |
| 6 | 5 | 3674 | 22650 | 12932 |
| 7 | 6 | 2676 | 11635 | 11306 |
| 8 | 7 | 1978 | 9218 | 10353 |
| 9 | 8 | 1898 | 16004 | 8556 |
| 10 | 9 | 1839 | 15345 | 6243 |
| 11 | 10 | 1611 | 11803 | 9006 |
| 12 | 11 | 3066 | 10753 | 13422 |
| 13 | 12 | 2569 | 12971 | 10634 |
| 14 | 13 | 4154 | 11870 | 13121 |
| 15 | 14 | 2078 | 25270 | 12218 |
| 16 | 15 | 2559 | 14015 | 17298 |
| 17 | 16 | 2031 | 16454 | 11803 |
| 18 | 17 | 6763 | 18305 | 8532 |
| 19 | 18 | 3200 | 14445 | 10950 |
| 20 | 19 | 3088 | 12103 | 8584 |
| 21 | 20 | 1901 | 14814 | 8367 |
| 22 | 21 | 3009 | 14006 | 8522 |
| 23 | 22 | 3214 | 10679 | 8825 |
| 24 | 23 | 1949 | 19293 | 10745 |

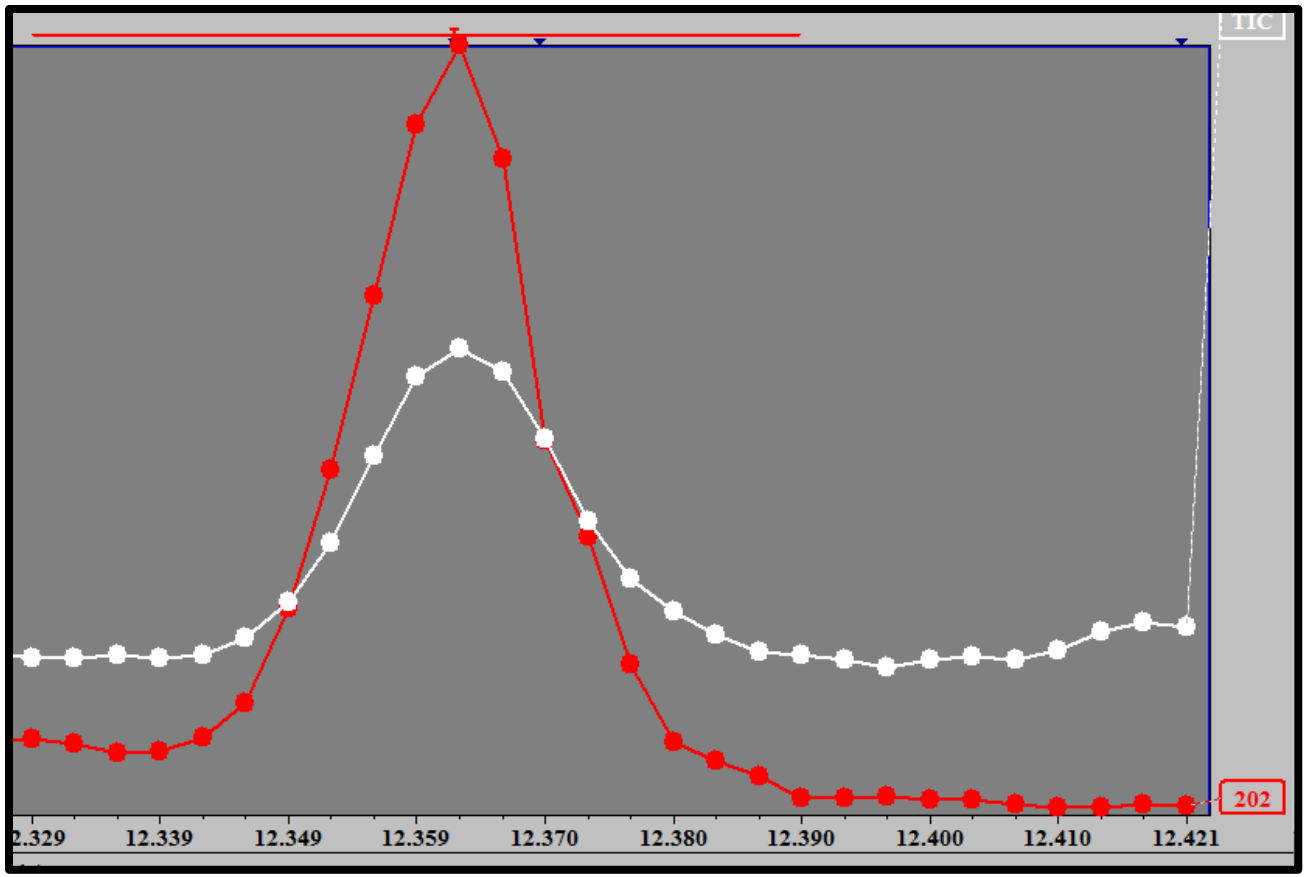
# Example XCMS Code in R

**Data processing aligns m/z values across many samples**



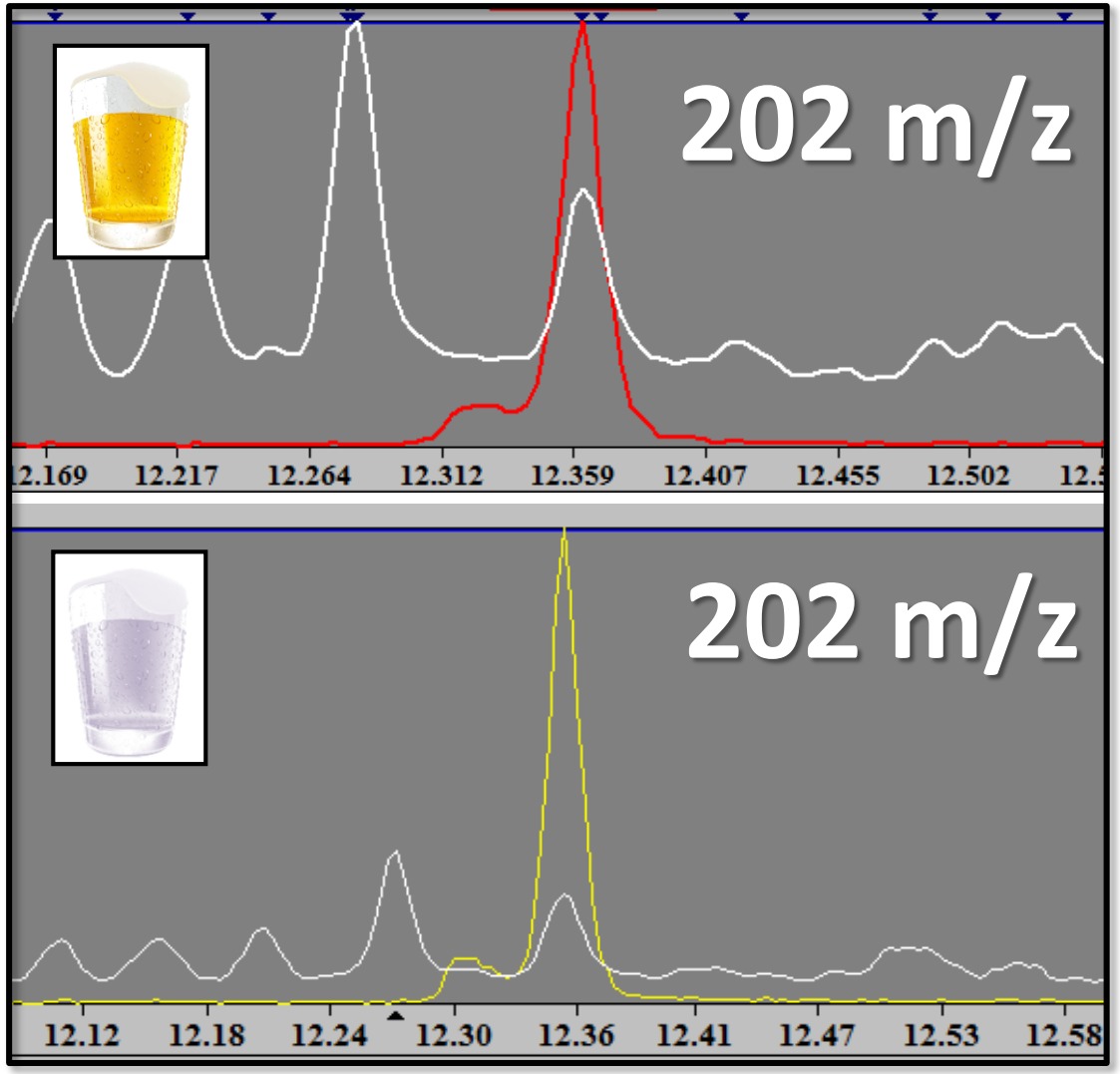

```
#LOAD XCMS
library(xcms)

#VIEW YOUR DATA
f1 <- xcmsRaw(dataset[1], profstep=0.5)
plotSurf(f1, log = FALSE, aspect = c(1, 1, .5), mzrange=c(50,610),
rtrange=c(300,1080))
```

# Step A: Detect Peaks



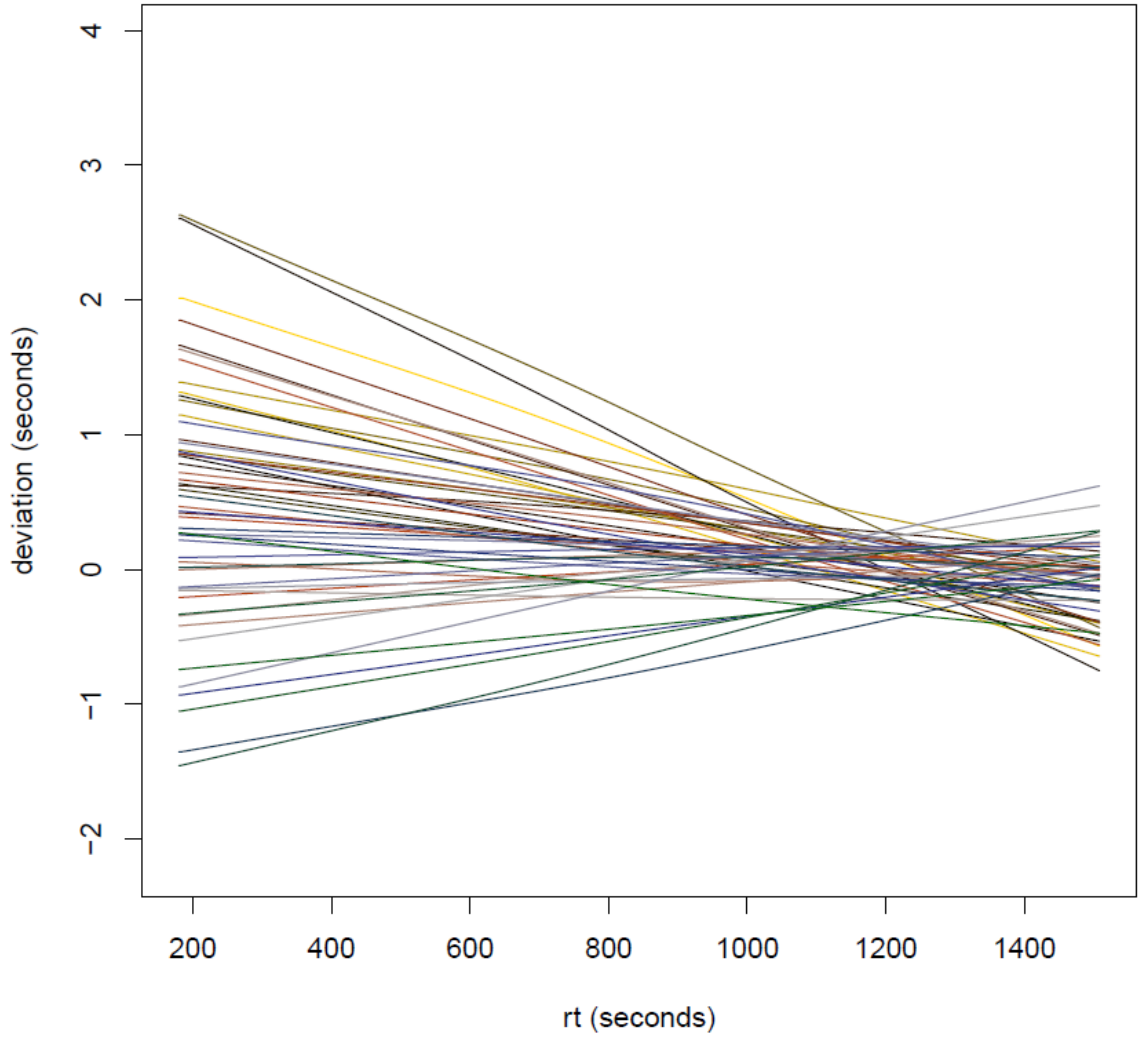

```
#LOAD XCMS
library(xcms)

#VIEW YOUR DATA
f1<-xcmsRaw(dataset[1], profstep=0.5)
plotSurf(f1, log = FALSE, aspect = c(1, 1, .5), mzrange=c(50,610),
rtrange=c(300,1080))

#DETECT PEAKS
xset <- xcmsSet(dataset, phenoData=dataset, nSlaves=4,
method = "matchedFilter", fwhm = 8, max = 500, snthresh = 5,
step = 0.1, steps = 2, mzdiff = 0.3, index = FALSE, sleep = 0)
xset

#GROUP PEAKS
xset <- group(xset, bw=2, minfrac=0.5, max = 100, mzwid=0.02)
xset

#RT CORRECTION
xset <- retcor(xset,  method="loess", family = "gaussian",
plottype = "mdevden", span=2, missing=2)
xset

#REGROUP
xset <- group(xset, bw=1.5, minfrac=0.45, max= 100, mzwid=0.04)
xset

#FILL PEAKS
xset <- fillPeaks.chrom(xset)
xset

##SAVE THE R OBJECT
save(xset, file="xset.Rdata")

#EXTRACT AND NORMALIZE DATA
xset2 <- groupval(xset,value="into")
xset3<-t(xset2)
TIC<-rowSums(xset3)
NORM<-1000000*xset3/TIC
xset4<-NORM
seq<-read.csv(file="seq.csv", header=TRUE, row.names=2)
xset5<-merge(seq, xset4, by="row.names", all.x=FALSE, all.y=FALSE)

#EXPORT DATA MATRIX
write.csv(xset5, file="veggie_data.csv", row.names=FALSE)

#DECONVOLUTION
library(RAMClustR)
RC<-ramclustR(xset, blocksize=500, st=1, sr=0.5)
str(RC)
```

# Step B: Group Peaks




```
#LOAD XCMS
library(xcms)

#VIEW YOUR DATA
f1<-xcmsRaw(dataset[1], profstep=0.5)
plotSurf(f1, log = FALSE, aspect = c(1, 1, .5), mzrange=c(50,610),
rtrange=c(300,1080))

#DETECT PEAKS
xset <- xcmsSet(dataset, phenoData=dataset, nSlaves=4,
method = "matchedFilter", fwhm = 8, max = 500, snthresh = 5,
step = 0.1, steps = 2, mzdiff = 0.3, index = FALSE, sleep = 0)
xset

#GROUP PEAKS
xset <- group(xset, bw=2, minfrac=0.5, max = 100, mzwid=0.02)
xset

#RT CORRECTION
xset <- retcor(xset,  method="loess", family = "gaussian",
plottype = "mdevden", span=2, missing=2)
xset

#REGROUP
xset <- group(xset, bw=1.5, minfrac=0.45, max= 100, mzwid=0.04)
xset

#FILL PEAKS
xset <- fillPeaks.chrom(xset)
xset

##SAVE THE R OBJECT
save(xset, file="xset.Rdata")

#EXTRACT AND NORMALIZE DATA
xset2 <- groupval(xset,value="into")
xset3<-t(xset2)
TIC<-rowSums(xset3)
NORM<-1000000*xset3/TIC
xset4<-NORM
seq<-read.csv(file="seq.csv", header=TRUE, row.names=2)
xset5<-merge(seq, xset4, by="row.names", all.x=FALSE, all.y=FALSE)

#EXPORT DATA MATRIX
write.csv(xset5, file="veggie_data.csv", row.names=FALSE)

#DECONVOLUTION
library(RAMClustR)
RC<-ramclustR(xset, blocksize=500, st=1, sr=0.5)
str(RC)
```

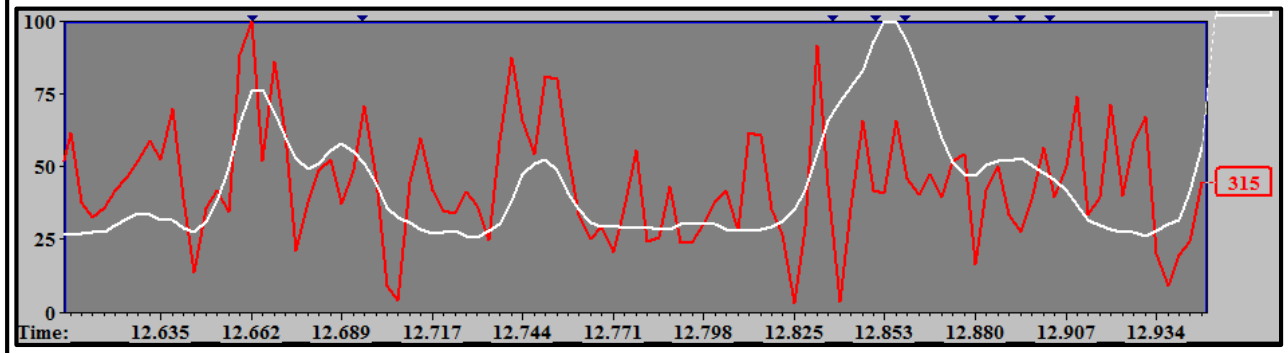
# Step C: RT Correction




```
#LOAD XCMS
library(xcms)

#VIEW YOUR DATA
f1<-xcmsRaw(dataset[1], profstep=0.5)
plotSurf(f1, log = FALSE, aspect = c(1, 1, .5), mzrange=c(50,610),
rtrange=c(300,1080))

#DETECT PEAKS
xset <- xcmsSet(dataset, phenoData=dataset, nSlaves=4,
method = "matchedFilter", fwhm = 8, max = 500, snthresh = 5,
step = 0.1, steps = 2, mzdiff = 0.3, index = FALSE, sleep = 0)
xset

#GROUP PEAKS
xset <- group(xset, bw=2, minfrac=0.5, max = 100, mzwid=0.02)
xset

#RT CORRECTION
xset <- retcor(xset,  method="loess", family = "gaussian",
plottype = "mdevden", span=2, missing=2)
xset

#REGROUP
xset <- group(xset, bw=1.5, minfrac=0.45, max= 100, mzwid=0.04)
xset

#FILL PEAKS
xset <- fillPeaks.chrom(xset)
xset

##SAVE THE R OBJECT
save(xset, file="xset.Rdata")

#EXTRACT AND NORMALIZE DATA
xset2 <- groupval(xset,value="into")
xset3<-t(xset2)
TIC<-rowSums(xset3)
NORM<-1000000*xset3/TIC
xset4<-NORM
seq<-read.csv(file="seq.csv", header=TRUE, row.names=2)
xset5<-merge(seq, xset4, by="row.names", all.x=FALSE, all.y=FALSE)

#EXPORT DATA MATRIX
write.csv(xset5, file="veggie_data.csv", row.names=FALSE)

#DECONVOLUTION
library(RAMClustR)
RC<-ramclustR(xset, blocksize=500, st=1, sr=0.5)
str(RC)
```

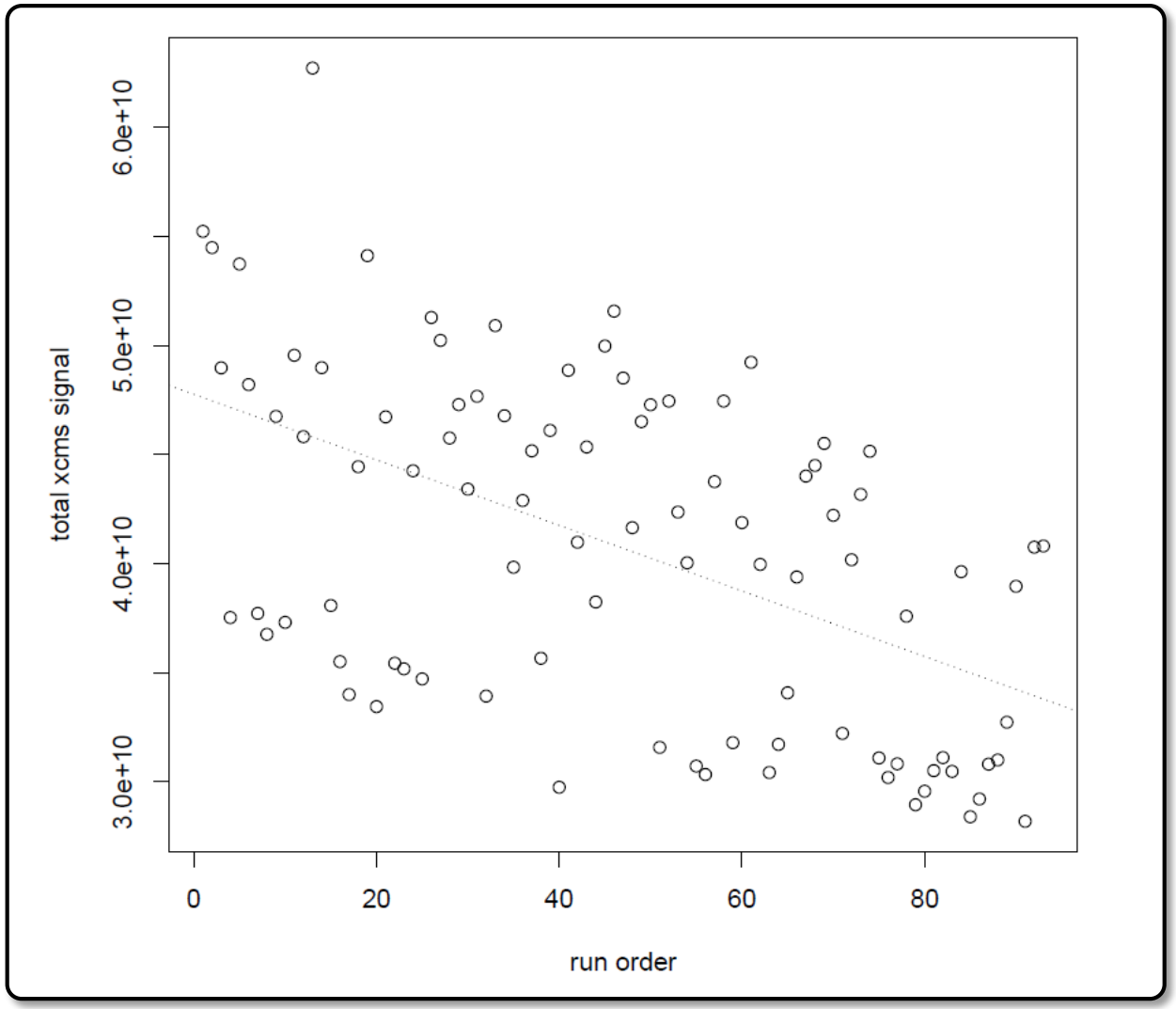
# Step D: Fill Peaks & Normalize

**‘0’ values are replaced with an estimation of local noise**




```
#LOAD XCMS
library(xcms)

#VIEW YOUR DATA
f1<-xcmsRaw(dataset[1], profstep=0.5)
plotSurf(f1, log = FALSE, aspect = c(1, 1, .5), mzrange=c(50,610),
rtrange=c(300,1080))

#DETECT PEAKS
xset <- xcmsSet(dataset, phenoData=dataset, nSlaves=4,
method = "matchedFilter", fwhm = 8, max = 500, snthresh = 5,
step = 0.1, steps = 2, mzdiff = 0.3, index = FALSE, sleep = 0)
xset

#GROUP PEAKS
xset <- group(xset, bw=2, minfrac=0.5, max = 100, mzwid=0.02)
xset

#RT CORRECTION
xset <- retcor(xset,  method="loess", family = "gaussian",
plottype = "mdevden", span=2, missing=2)
xset

#REGROUP
xset <- group(xset, bw=1.5, minfrac=0.45, max= 100, mzwid=0.04)
xset

#FILL PEAKS
xset <- fillPeaks.chrom(xset)
xset

##SAVE THE R OBJECT
save(xset, file="xset.Rdata")

#EXTRACT AND NORMALIZE DATA
xset2 <- groupval(xset,value="into")
xset3<-t(xset2)
TIC<-rowSums(xset3)
NORM<-1000000*xset3/TIC
xset4<-NORM
seq<-read.csv(file="seq.csv", header=TRUE, row.names=2)
xset5<-merge(seq, xset4, by="row.names", all.x=FALSE, all.y=FALSE)

#EXPORT DATA MATRIX
write.csv(xset5, file="veggie_data.csv", row.names=FALSE)

#DECONVOLUTION
library(RAMClustR)
RC<-ramclustR(xset, blocksize=500, st=1, sr=0.5)
str(RC)
```

# Step D: Fill Peaks & Normalize




```
#LOAD XCMS
library(xcms)

#VIEW YOUR DATA
f1<-xcmsRaw(dataset[1], profstep=0.5)
plotSurf(f1, log = FALSE, aspect = c(1, 1, .5), mzrange=c(50,610),
rtrange=c(300,1080))

#DETECT PEAKS
xset <- xcmsSet(dataset, phenoData=dataset, nSlaves=4,
method = "matchedFilter", fwhm = 8, max = 500, snthresh = 5,
step = 0.1, steps = 2, mzdiff = 0.3, index = FALSE, sleep = 0)
xset

#GROUP PEAKS
xset <- group(xset, bw=2, minfrac=0.5, max = 100, mzwid=0.02)
xset

#RT CORRECTION
xset <- retcor(xset,  method="loess", family = "gaussian",
plottype = "mdevden", span=2, missing=2)
xset

#REGROUP
xset <- group(xset, bw=1.5, minfrac=0.45, max= 100, mzwid=0.04)
xset

#FILL PEAKS
xset <- fillPeaks.chrom(xset)
xset

##SAVE THE R OBJECT
save(xset, file="xset.Rdata")

#EXTRACT AND NORMALIZE DATA
xset2 <- groupval(xset,value="into")
xset3<-t(xset2)
TIC<-rowSums(xset3)
NORM<-1000000*xset3/TIC
xset4<-NORM
seq<-read.csv(file="seq.csv", header=TRUE, row.names=2)
xset5<-merge(seq, xset4, by="row.names", all.x=FALSE, all.y=FALSE)

#EXPORT DATA MATRIX
write.csv(xset5, file="veggie_data.csv", row.names=FALSE)

#DECONVOLUTION
library(RAMClustR)
RC<-ramclustR(xset, blocksize=500, st=1, sr=0.5)
str(RC)
```

# The End Product Is a Data Matrix of Samples and Molecular Features

***XCMS aligned m/z and retention time among all samples***

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | sample | 76.8759_227.011 | 51.1409_227.876 | 78.1372_228.3 | 68.9466_228.573 | 64.056_228.733 | 125.9805_ | 56.9305_2 | 60.7199_2 | 55.163_22 | 117.7181_ |
| 2 | 1 | 5049 | 15766 | 15085 | 3477 | 419 | 1381 | 2079 | 2459 | 1879 | 16972 |
| 3 | 2 | 1913 | 18305 | 13923 | 4910 | 656 | 1436 | 1992 | 3166 | 1517 | 12409 |
| 4 | 3 | 3364 | 21773 | 14991 | 4361 | 320 | 1306 | 2713 | 2687 | 3503 | 10642 |
| 5 | 4 | 7636 | 11918 | 9107 | 4913 | 555 | 1541 | 1435 | 2456 | 2235 | 14140 |
| 6 | 5 | 3674 | 22650 | 12932 | 3626 | 220 | 1004 | 2892 | 2366 | 788 | 10578 |
| 7 | 6 | 2676 | 11635 | 11306 | 4973 | 211 | 1469 | 1479 | 1825 | 1004 | 8187 |
| 8 | 7 | 1978 | 9218 | 10353 | 7043 | 130 | 1588 | 1370 | 2343 | 1360 | 8170 |
| 9 | 8 | 1898 | 16004 | 8556 | 3950 | 412 | 1174 | 659 | 2044 | 3187 | 10840 |
| 10 | 9 | 1839 | 15345 | 6243 | 5337 | 1301 | 1395 | 1179 | 2337 | 309 | 9708 |
| 11 | 10 | 1611 | 11803 | 9006 | 5943 | 910 | 1185 | 1309 | 2048 | 507 | 9740 |
| 12 | 11 | 3066 | 10753 | 13422 | 6014 | 1024 | 1173 | 2248 | 2869 | 2261 | 9911 |
| 13 | 12 | 2569 | 12971 | 10634 | 6678 | 931 | 1661 | 919 | 1484 | 2243 | 7538 |
| 14 | 13 | 4154 | 11870 | 13121 | 3040 | 635 | 1541 | 889 | 3437 | 1327 | 8443 |
| 15 | 14 | 2078 | 25270 | 12218 | 5695 | 174 | 1228 | 511 | 2131 | 3319 | 12868 |
| 16 | 15 | 2559 | 14015 | 17298 | 2299 | 160 | 1847 | 331 | 2939 | 1088 | 9282 |
| 17 | 16 | 2031 | 16454 | 11803 | 5422 | 678 | 1057 | 764 | 1339 | 631 | 7523 |
| 18 | 17 | 6763 | 18305 | 8532 | 4020 | 1286 | 1195 | 3230 | 2506 | 1428 | 11523 |
| 19 | 18 | 3200 | 14445 | 10950 | 4053 | 684 | 1562 | 2054 | 1719 | 1792 | 12295 |
| 20 | 19 | 3088 | 12103 | 8584 | 4427 | 1756 | 1321 | 4106 | 2407 | 1215 | 8350 |
| 21 | 20 | 1901 | 14814 | 8367 | 2880 | 309 | 1211 | 498 | 1947 | 1689 | 10607 |
| 22 | 21 | 3009 | 14006 | 8522 | 1877 | 131 | 946 | 1764 | 2074 | 1587 | 11608 |
| 23 | 22 | 3214 | 10679 | 8825 | 3600 | 85 | 1344 | 2439 | 1573 | 2313 | 15287 |
| 24 | 23 | 1949 | 19293 | 10745 | 5730 | 1330 | 609 | 1917 | 1787 | 1529 | 7648 |

# Deconvolution Reduces Data Complexity and Improves Quantitation and Statistics

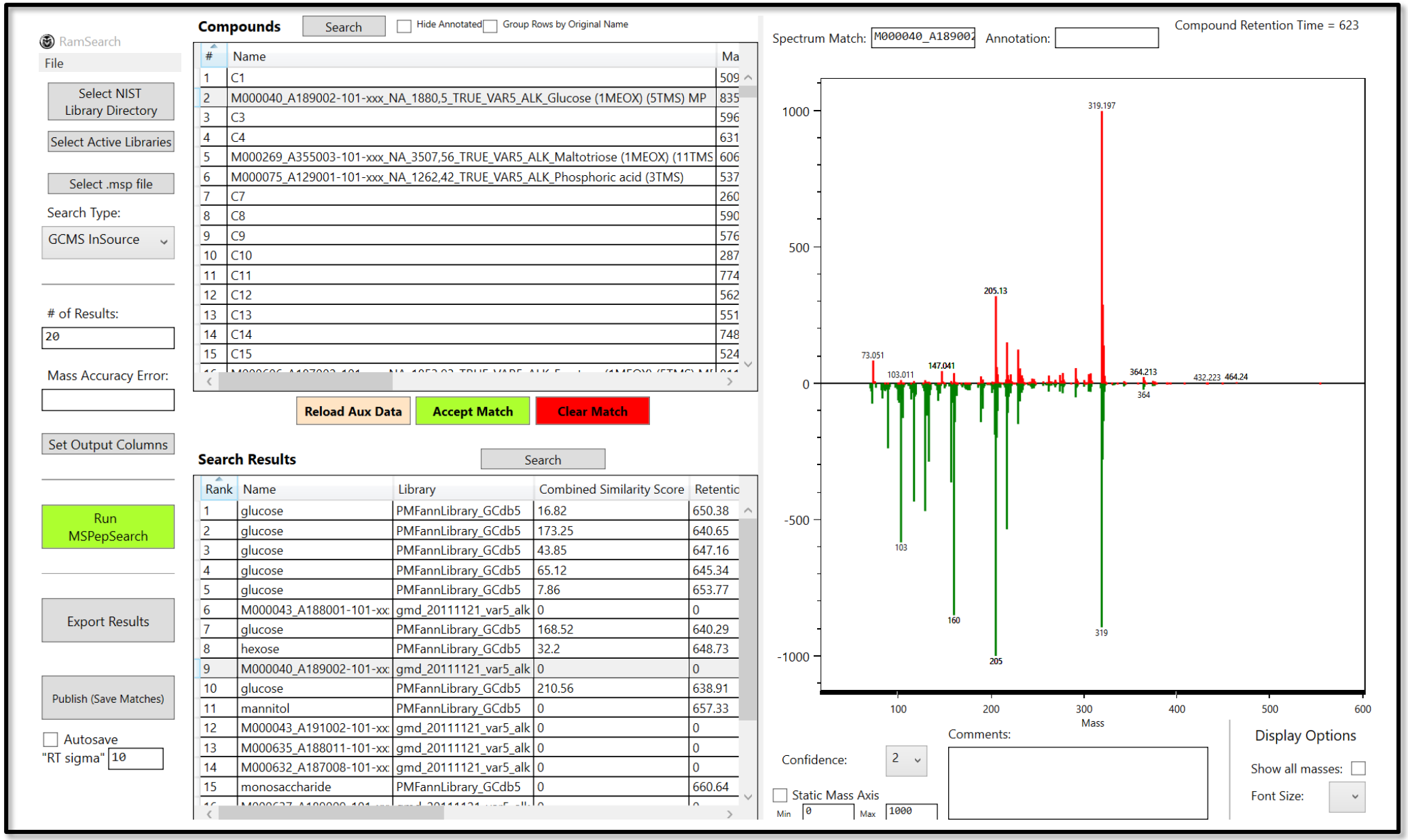
| | A | B | C |
|---|---|---|---|
| 1 | sample | 76.8759_227.011 | 51.1409_227.876 |
| 2 | 1 | 5049 | 15766 |
| 3 | 2 | 1913 | 18305 |
| 4 | 3 | 3364 | 21773 |
| 5 | 4 | 7636 | 11918 |
| 6 | 5 | 3674 | 22650 |
| 7 | 6 | 2676 | 11635 |
| 8 | 7 | 1978 | 9218 |
| 9 | 8 | 1898 | 16004 |
| 10 | 9 | 1839 | 15345 |
| 11 | 10 | 1611 | 11803 |
| 12 | 11 | 3066 | 10753 |
| 13 | 12 | 2569 | 12971 |
| 14 | 13 | 4154 | 11870 |
| 15 | 14 | 2078 | 25270 |
| 16 | 15 | 2559 | 14015 |
| 17 | 16 | 2031 | 16454 |
| 18 | 17 | 6763 | 18305 |
| 19 | 18 | 3200 | 14445 |
| 20 | 19 | 3088 | 12103 |
| 21 | 20 | 1901 | 14814 |
| 22 | 21 | 3009 | 14006 |
| 23 | 22 | 3214 | 10679 |
| 24 | 23 | 1949 | 19293 |

**RAMClust deconvolution algorithm**

```
Name: C1
SYNON: $:00in-source
SYNON: $:04
SYNON: $:05NA
SYNON: $:06QUAD
SYNON: $:07Thermo ISQ
SYNON: $:09Thermo Trace GC: TG-5MS column
SYNON: $:10EI
SYNON: $:11P
SYNON: $:12NA
SYNON: $:1450-650
SYNON: $:16NA
Comment: Rt=894.58  Contributor=Heuberger  Study=Beer
Num Peaks: 226
361.142 43898864 362.1982 14306227 129.0275 9795326 21
451.1973 2617899 438.2199 2326944 131.0569 2225511 169
 229.0679 1279031 439.2148 1249159 320.1808 1204107 19
 774139 273.1298 722619 204.1072 696785 305.132 677224
511265 149.0883 485191 365.1848 467293 332.1928 448628
330261 232.1325 329233 241.0904 329101 306.1674 320391
263433 246.1461 262180 215.0734 234777 227.105 212544
145653 346.1988 145649 454.2283 143148 220.1388 138744
120659 207.0822 115630 183.0561 107308 234.1325 103054
 141.0658 84835 175.0647 84744 263.1023 84691 177.0717
 67901 374.1817 65978 134.1186 63698 349.1869 57906 37
278.1496 55147 150.1563 54756 308.1728 53185 161.0897
 41059 249.1223 40932 455.2259 40475 151.0812 39521 18
379.1642 32610 287.1217 31666 310.1953 30767 433.2124
21731 355.2281 21589 172.1683 21492 381.1674 20880 208
17458 315.1486 17348 184.1429 17300 224.1193 16361 336
213.0978 13304 164.1691 13297 178.1456 13128 237.0965
 9552 337.1601 9251 165.1346 8896 431.166 8282 469.203
 250.1405 7044 382.1824 6883 351.1818 6728 422.2081 66
 4440 383.1604 4141 137.0644 4030 427.1343 3942 400.88
```

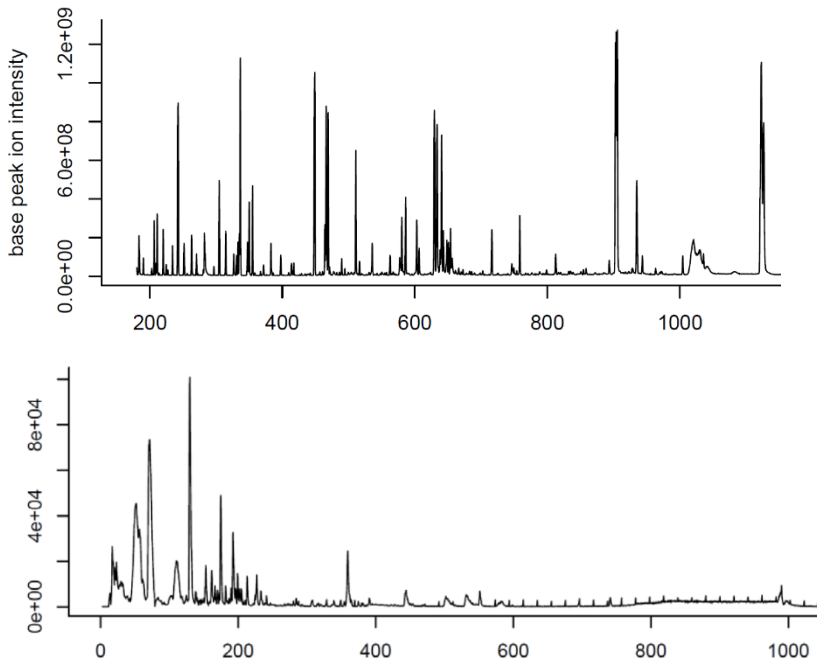
| sample | C1 | C2 | C3 |
|---|---|---|---|
| 1 | 3082 | 2686 | 104967 |
| 2 | 1662 | 1987 | 91740 |
| 3 | 1635 | 2522 | 84272 |
| 4 | 2544 | 1906 | 78988 |
| 5 | 3155 | 2083 | 72516 |
| 6 | 3004 | 2168 | 98226 |
| 7 | 2471 | 1836 | 93222 |
| 8 | 2241 | 2042 | 94619 |
| 9 | 1742 | 1950 | 83389 |
| 10 | 2691 | 2311 | 100043 |
| 11 | 1745 | 2352 | 82148 |
| 12 | 2253 | 2560 | 84004 |
| 13 | 1502 | 2166 | 65728 |
| 14 | 2955 | 2256 | 95223 |
| 15 | 2381 | 2304 | 90304 |
| 16 | 1918 | 2117 | 82843 |
| 17 | 2346 | 2575 | 101296 |
| 18 | 2103 | 2829 | 110123 |
| 19 | 1696 | 2162 | 88921 |
| 20 | 1982 | 1656 | 95223 |
| 21 | 1825 | 2376 | 101436 |
| 22 | 1945 | 2410 | 91912 |
| 23 | 1941 | 2192 | 85924 |
| 24 | 1770 | 1608 | 73445 |
| 25 | 2581 | 2393 | 92053 |

# Metabolite Annotation Using RamSearch

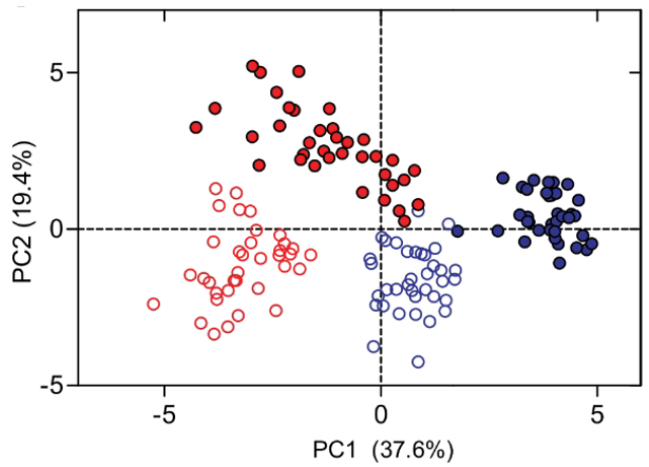


# Data Analysis Involves Many Types of Statistics

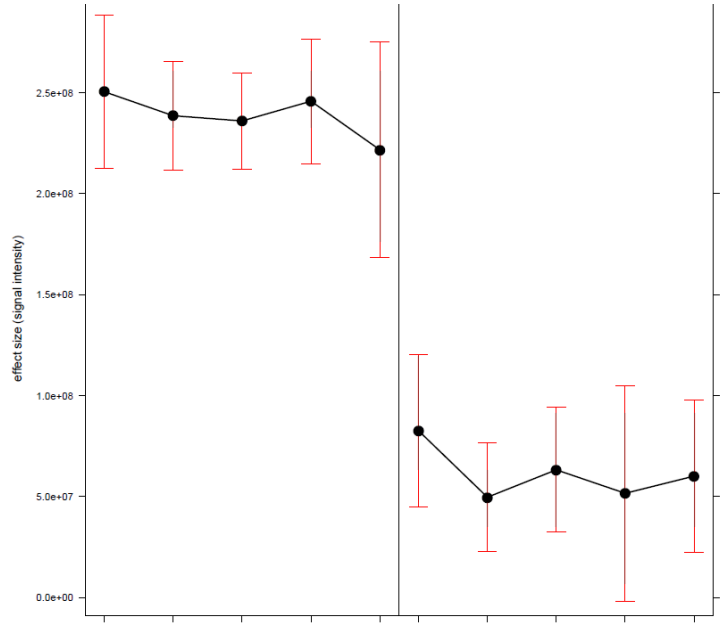
## Quality Control Statistics




## Overview Statistics

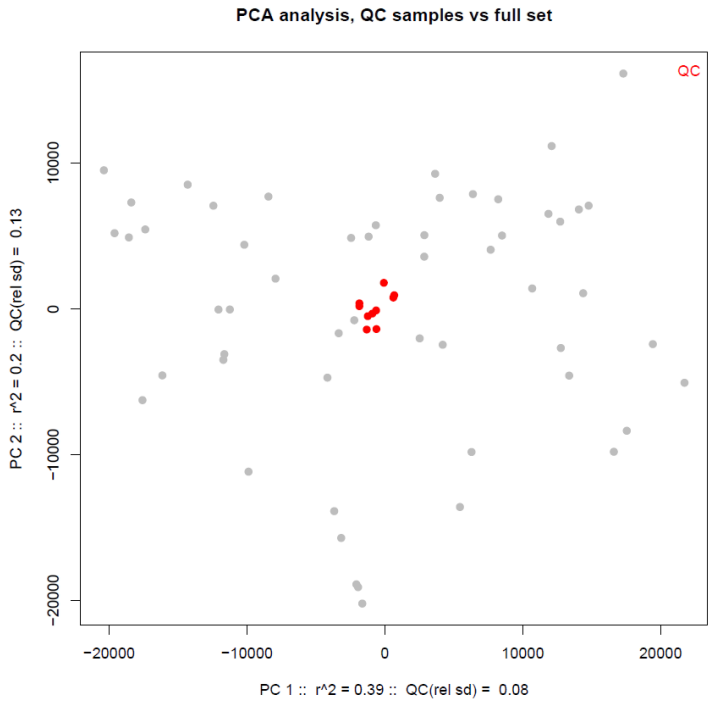



## Comparative Statistics

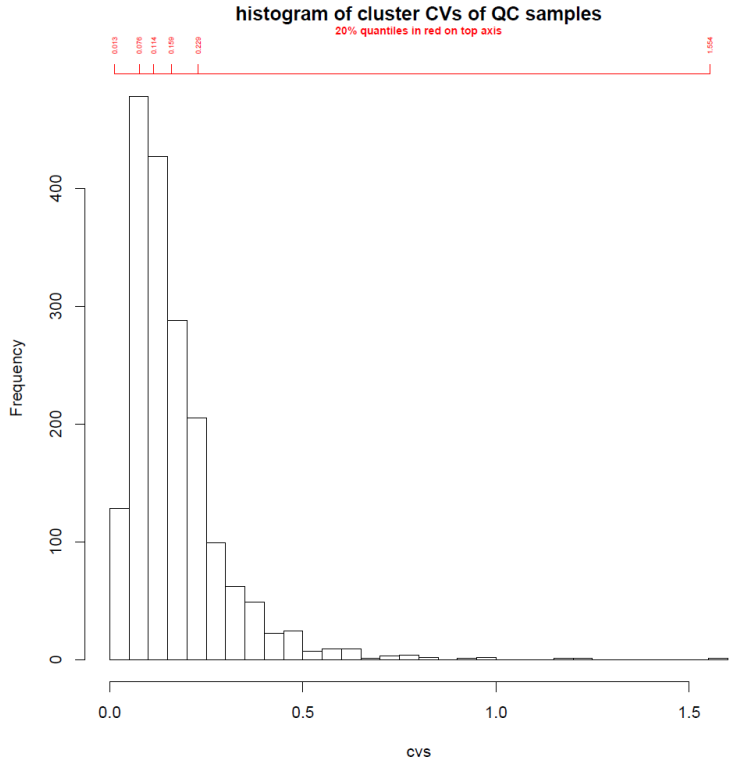


# Quality Control/Assurance Statistics

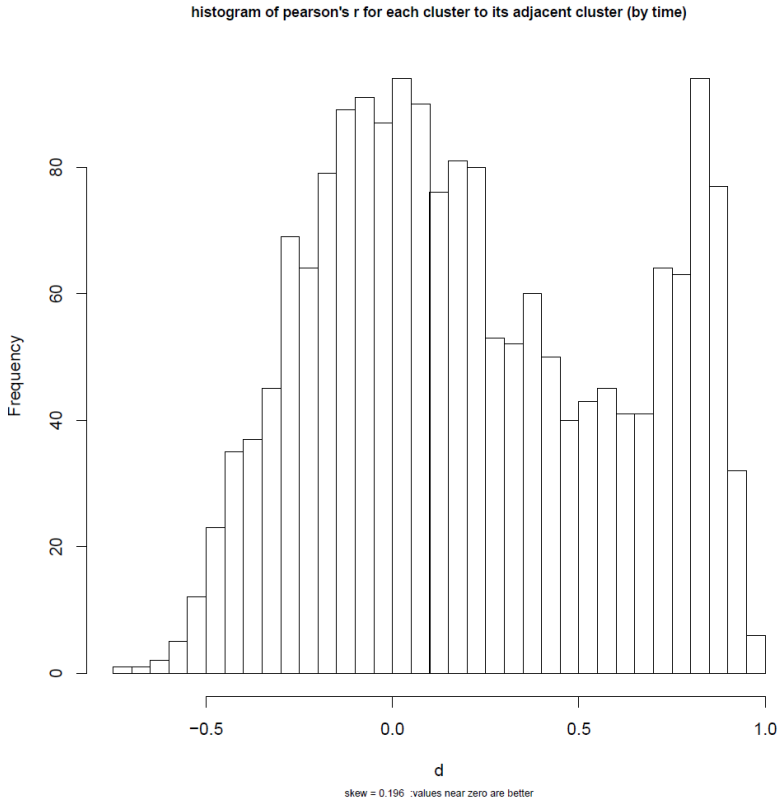
## PCA of pooled QC samples




## CVs of QC samples

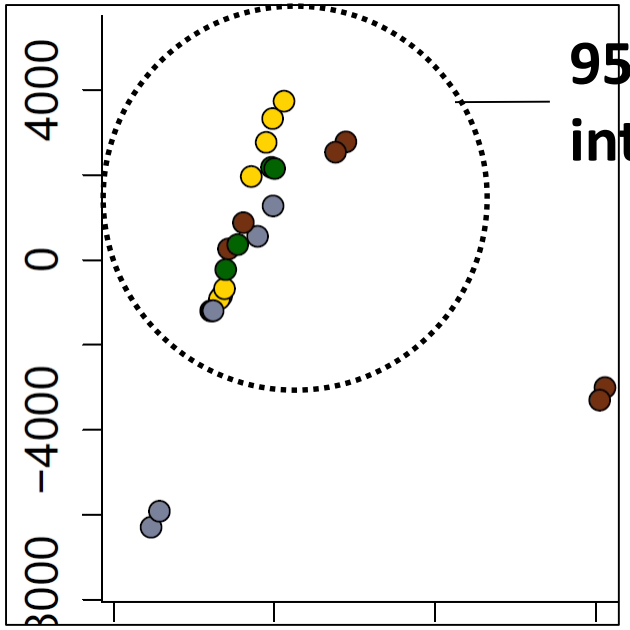



## Deconvolution QA



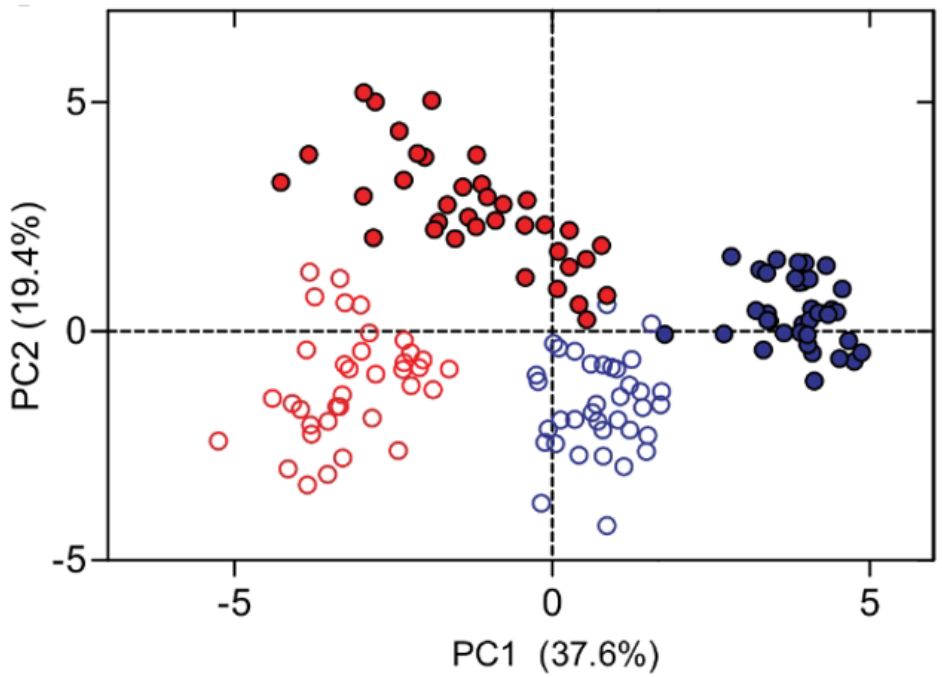
# PCA Provides A Great Overview of Metabolite Data

## PCA can detect outliers




95% confidence interval

## PCA can show trends in chemical variation



# ANOVA and Regression Are Important Univariate Models (Use FDR Correction!)

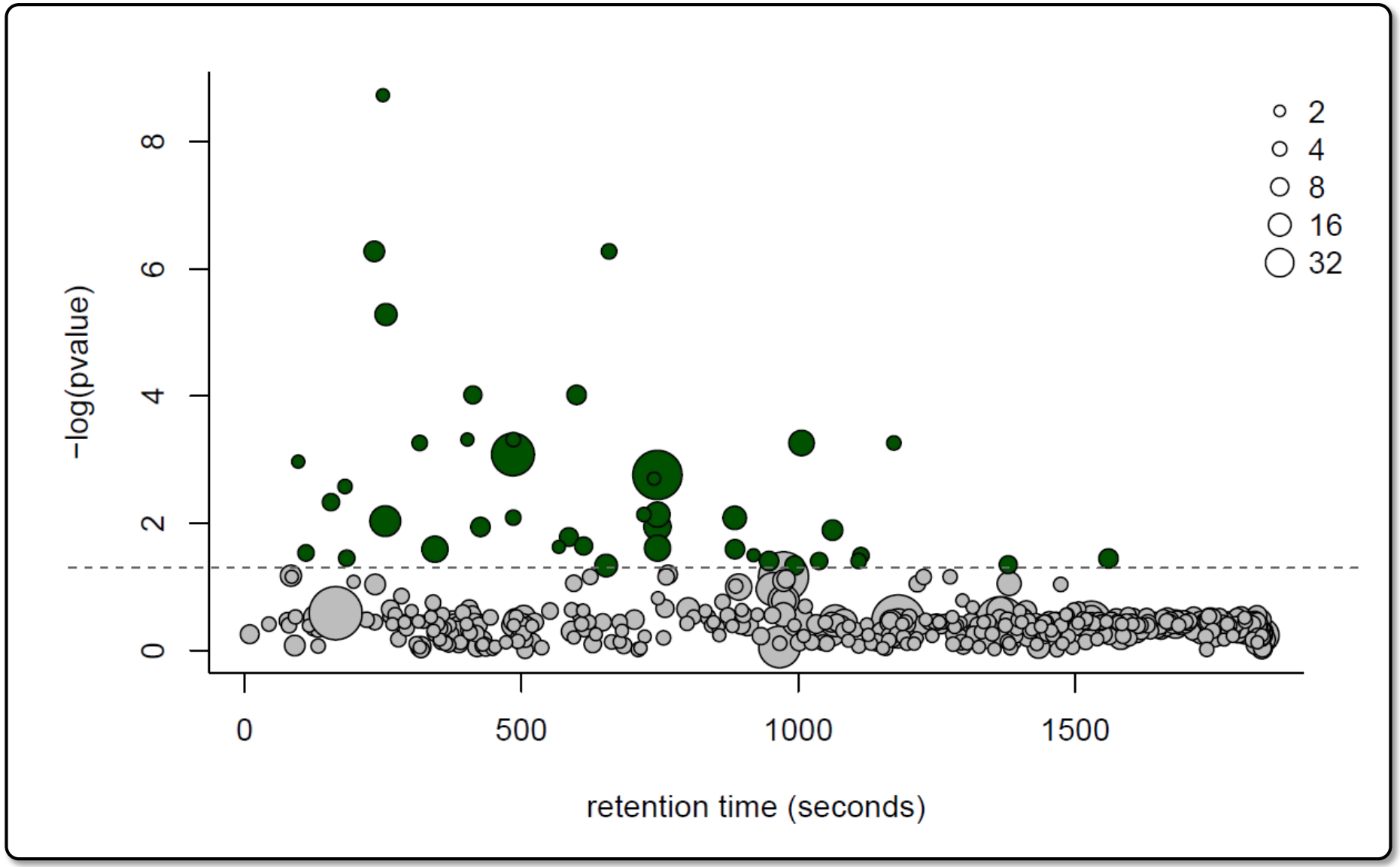
**met ~ type + time + type*time**

**time = β (met) + intercept = 0?**

***each metabolite is assigned a p-value for each test***

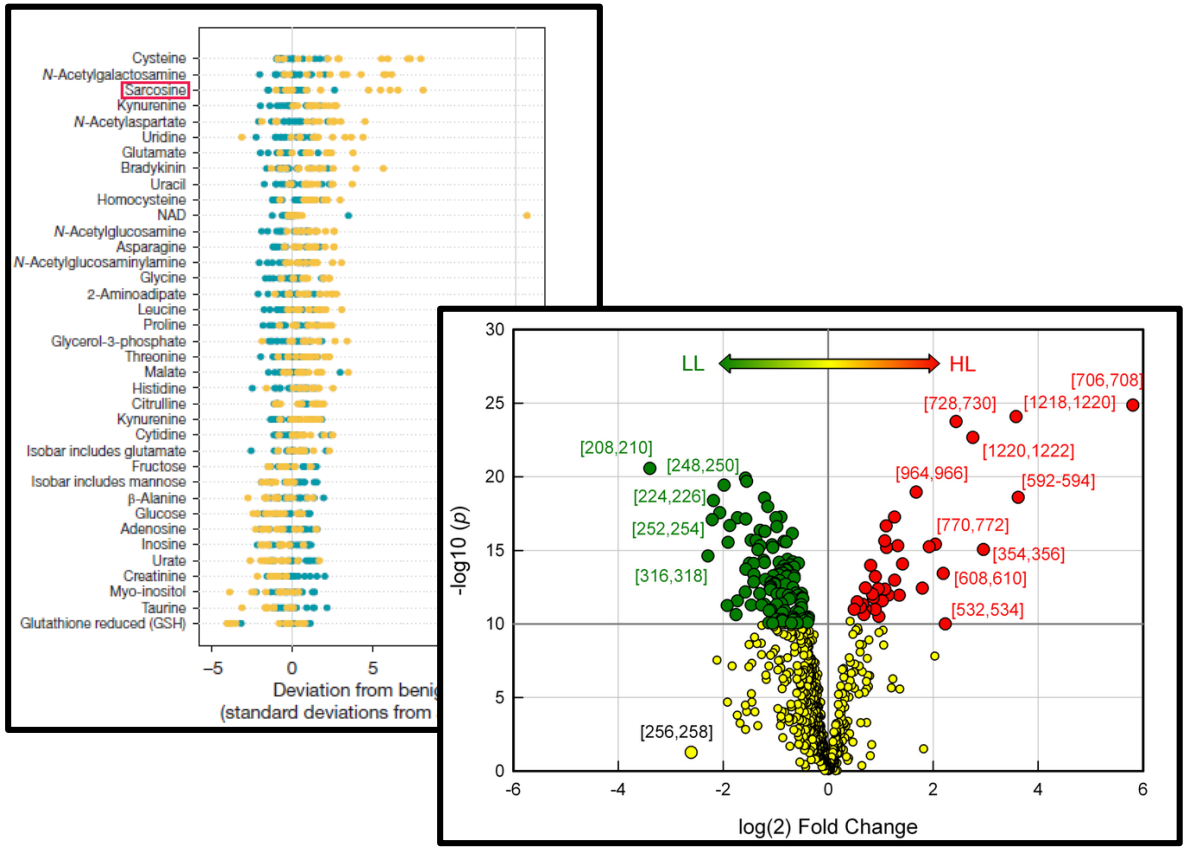
| 1 | sample | maltose | tryptophai | citric acid | lactic acid | glucose |
|---|---|---|---|---|---|---|
| 2 | NB - time 1 - rep 1 | 3082 | 2686 | 104967 | 4430 | 192 |
| 3 | NB - time 1 - rep 2 | 1662 | 1987 | 91740 | 2849 | 382 |
| 4 | NB - time 1 - rep 3 | 1635 | 2522 | 84272 | 4288 | 165 |
| 5 | Odell - time 1 - rep 1 | 2544 | 1906 | 78988 | 3678 | 136 |
| 6 | Odell - time 1 - rep 2 | 3155 | 2083 | 72516 | 3496 | 309 |
| 7 | Odell - time 1 - rep 3 | 3004 | 2168 | 98226 | 4148 | 263 |
| 8 | NB - time 2 - rep 1 | 2471 | 1836 | 93222 | 5716 | 155 |
| 9 | NB - time 2 - rep 2 | 2241 | 2042 | 94619 | 5780 | 182 |
| 10 | NB - time 2 - rep 3 | 1742 | 1950 | 83389 | 6284 | 157 |
| 11 | Odell - time 2 - rep 1 | 2691 | 2311 | 100043 | 3837 | 211 |
| 12 | Odell - time 2 - rep 2 | 1745 | 2352 | 82148 | 3456 | 228 |
| 13 | Odell - time 2 - rep 3 | 2253 | 2560 | 84004 | 4212 | 110 |
| 14 | NB - time 3 - rep 1 | 1502 | 2166 | 65728 | 4613 | 459 |
| 15 | NB - time 3 - rep 2 | 2955 | 2256 | 95223 | 3570 | 219 |
| 16 | NB - time 3 - rep 3 | 2381 | 2304 | 90304 | 3141 | 159 |
| 17 | Odell - time 3 - rep 1 | 1918 | 2117 | 82843 | 2460 | 218 |
| 18 | Odell - time 3 - rep 2 | 2346 | 2575 | 101296 | 4915 | 218 |
| 19 | Odell - time 3 - rep 3 | 2103 | 2829 | 110123 | 2985 | 429 |

# P-Value Bubble Plots (P values vs. Retention Time)



# There Are Many Ways To View And Interpret Data (The Statistics Toolbox)

- z scores
- hierarchical clustering
- heat mapping
- fold differences
- multivariate regressions (e.g. PLS)
  - discriminant analysis (A vs. B)
  - regression: (y ~ met1 + met2….)
- data integration methods (O2PLS)
- networking

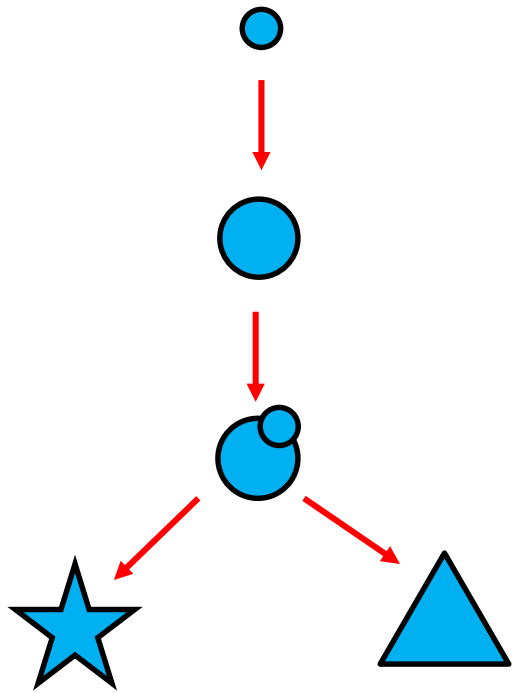



Vol 457|12 February 2009|doi:10.1038/nature07762

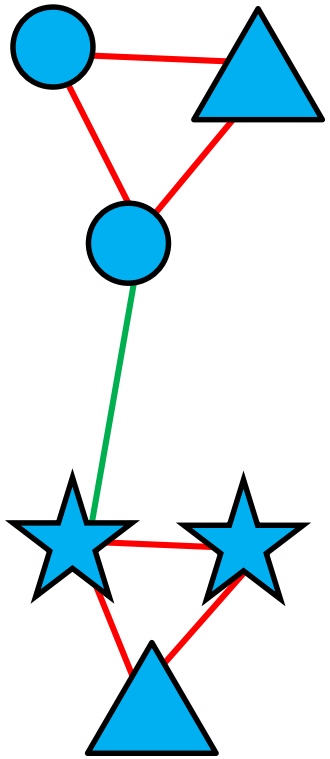
Grace, Stephen C., Stephen Embry, and Heng Luo. "Haystack, a web-based tool for metabolomics research." *BMC bioinformatics* 15.Suppl 11 (2014): S12.

# Organizing The Data Is An Important Step To Facilitate Interpretation

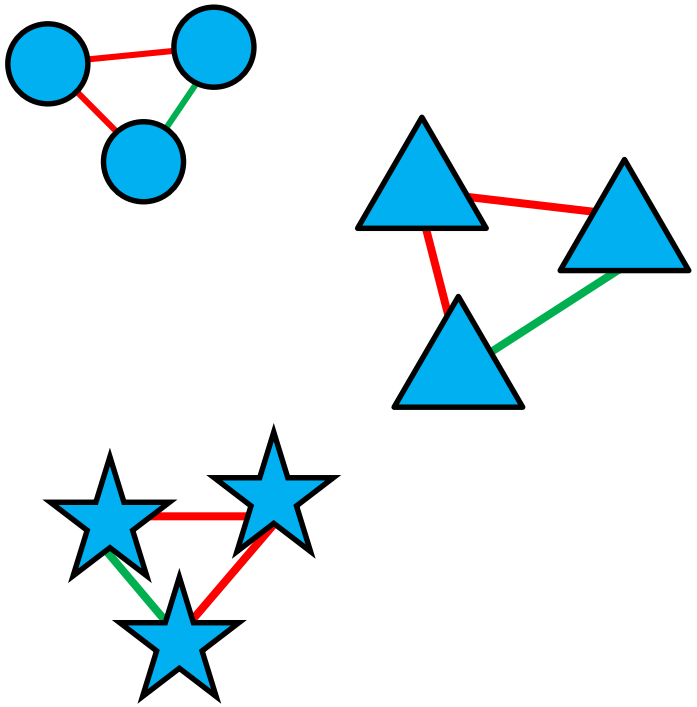
**Biochemistry**



**Co-variation**

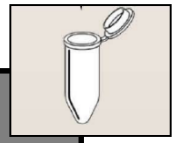


**Chemical Structures**

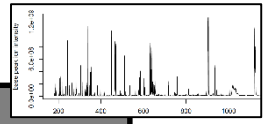


1 Develop a hypothesis

2 Extract metabolites

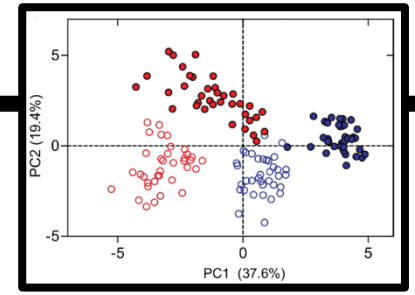


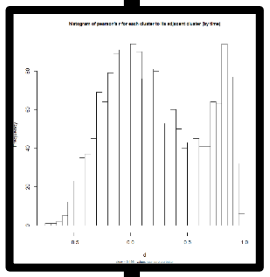
3 Detect metabolites



| | sample | maltose | tryptophai | citric acid | lactic acid | glucose |
|---|---|---|---|---|---|---|
| 2 | NB - time 1 - rep 1 | 3082 | 2686 | 104967 | 4430 | 192 |
| 3 | NB - time 1 - rep 2 | 1662 | 1987 | 91740 | 2849 | 382 |
| 4 | NB - time 1 - rep 3 | 1635 | 2522 | 84272 | 4288 | 165 |
| 5 | Odell - time 1 - rep 1 | 2544 | 1906 | 78988 | 3678 | 136 |
| 6 | Odell - time 1 - rep 2 | 3155 | 2083 | 72516 | 3496 | 305 |
| 7 | Odell - time 1 - rep 3 | 3004 | 2168 | 98226 | 4148 | 262 |
| 8 | NB - time 2 - rep 1 | 2471 | 1836 | 93222 | 5716 | 155 |
| 9 | NB - time 2 - rep 2 | 2241 | 2042 | 94619 | 5780 | 182 |
| 10 | NB - time 2 - rep 3 | 1742 | 1950 | 83389 | 6284 | 152 |
| 11 | Odell - time 2 - rep 1 | 2691 | 2311 | 100043 | 3837 | 211 |
| 12 | Odell - time 2 - rep 2 | 1745 | 2352 | 82148 | 3456 | 228 |
| 13 | Odell - time 2 - rep 3 | 2253 | 2560 | 84004 | 4212 | 110 |
| 14 | NB - time 3 - rep 1 | 1502 | 2166 | 65728 | 4613 | 459 |
| 15 | NB - time 3 - rep 2 | 2955 | 2256 | 95223 | 3570 | 219 |
| 16 | NB - time 3 - rep 3 | 2381 | 2304 | 90304 | 3141 | 159 |
| 17 | Odell - time 3 - rep 1 | 1918 | 2117 | 82843 | 2460 | 218 |
| 18 | Odell - time 3 - rep 2 | 2346 | 2575 | 101296 | 4915 | 218 |
| 19 | Odell - time 3 - rep 3 | 2103 | 2829 | 110123 | 2985 | 429 |

4 Process Data



5 Analyze Data

***A better understanding of chemical variation in your system!***

# Summary

- GC-MS is a robust method to detect non-volatile metabolites
- Metabolomics is a method that compares metabolites profiles, and this may be useful to investigate hypotheses in brewing science
- A metabolomics workflow consists of a extraction, detection, and analysis methods

**Resources**

- Metacyc (Pathway Tools)
- Gramene Metabolic Map for cereals
- Metlin (Scripps)
- Lipidmaps
- Human Metabolome Database
- Foodb.ca
- Metabolomics Society
- North American Metabolomics Society

Colorado State University